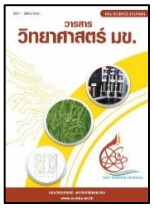
KKU SCIENCE JOURNAL
Journal Home Page : https://ph01.tci-thaijo.org/index.php/KKUSciJ
Published by the Faculty of Science, Khon Kaen University, Thailand



ปัจจัยที่ส่งผลต่อการตัดสินใจเข้าศึกษาต่อในระดับปริญญาตรี คณะบริหารธุรกิจและศิลปศาสตร์ โดยใช้เทคนิคการคัดเลือกคุณลักษณะและเทคนิคการทำเหมืองข้อมูล

Factors Affecting Decision to Study in Bachelor's Degree of Faculty of Business Administrator and Liberal Art Using Feature Selection and Data Mining Techniques

วรการ ใจดี[1] และ นรินทร์ จิวิตัน[1*]

Worakarn Jaidee[1] and Narin Jiwitan[1*]

[1]คณะบริหารธุรกิจและศิลปศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา จังหวัดเชียงใหม่ 50300

[1]Faculty of Business Administrator and Liberal Art, Rajamangala University of Technology Lanna Chiang Mai, Chiang Mai, 50300, Thailand

บทคัดย่อ

การวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาปัจจัยที่ส่งผลต่อการตัดสินใจเข้าศึกษาต่อในระดับปริญญาตรี คณะบริหารธุรกิจและศิลปศาสตร์ และเพื่อเปรียบเทียบประสิทธิภาพเทคนิคการทำเหมืองข้อมูล จากข้อมูลพื้นฐานการรับสมัครนักศึกษาของงานส่งเสริมวิชาการและงานทะเบียน มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา เชียงใหม่ ในช่วงปีการศึกษา 2563 – 2565 มีจำนวนข้อมูลผู้สมัครทั้งหมด 2,509 รายการ โดยคณะผู้วิจัยได้นำข้อมูลมาทำการวิเคราะห์ปัจจัยโดยใช้เทคนิคการคัดเลือกคุณลักษณะและเทคนิคการทำเหมืองข้อมูล โดยเทคนิคการคัดเลือกคุณลักษณะที่ใช้ประกอบไปด้วย 3 เทคนิค ได้แก่ 1) เทคนิคการหาค่าสถิติไคสแควร์ 2) เทคนิคการหาค่าสัมประสิทธิ์สหสัมพันธ์ และ 3) เทคนิคการหาค่าเกนความรู้ และเทคนิคการทำเหมืองข้อมูลที่ใช้ประกอบไปด้วย 6 เทคนิค ได้แก่ 1) เทคนิคต้นไม้ตัดสินใจ 2) เทคนิคป่าสุ่ม 3) เทคนิคเกรเดียนท์บูตทรีส์ 4) เทคนิคนาอีฟเบย์ 5) เทคนิคการถดถอยโลจีสติก และ 6) เทคนิคการโหวต ผลการศึกษาพบว่า ปัจจัยที่ส่งผลต่อการตัดสินใจเข้าศึกษาต่อในระดับปริญญาตรีมากที่สุด 3 อันดับแรกจากทุกเทคนิคของการคัดเลือกคุณลักษณะได้แก่ สาขาวิชาหรือหลักสูตรที่เลือก สายการเรียนเดิม และวุฒิการศึกษาเดิม ส่วนเทคนิคการทำเหมืองข้อมูลที่มีค่าความถูกต้องสูงสุดในการพยากรณ์คือเทคนิคการโหวต มีค่าความถูกต้องเท่ากับ 73.44% โดยมีค่าความถูกต้องสูงกว่าการใช้เทคนิคป่าสุ่ม เทคนิคการถดถอยโลจีสติก เทคนิคต้นไม้ตัดสินใจ เทคนิคนาอีฟเบย์ และเทคนิคเกรเดียนท์บูตทรีส์ ซึ่งมีค่าความถูกต้องเท่ากับ 71.45% 71.05% 68.26% 65.60% 64.81% ตามลำดับ

*Corresponding Author, E-mail: narin@rmutl.ac.th

Received: date: 30 May 2023 | Revised date: 4 July 2023 | Accepted date: 10 July 2023

ABSTRACT

The objective of this research was to study the factors affecting the decision to study at the bachelor's degree level, Faculty of Business Administration and Liberal Arts and to compare the effectiveness of data mining techniques. Based on the basic information on student recruitment for academic and registration of Rajamangala University of Technology Lanna, Chiang Mai, during the academic year 2020 - 2022, there were a total of 2,509 applicants. The research team has brought the data for factor analysis using feature selection and data mining techniques. The feature selection techniques used consisted of 3 techniques, namely 1) Chi Square Statistic, 2) Correlation Based, and 3) Information Gain. The data mining techniques used consisted of 6 techniques, namely 1) Decision Tree, 2) Random Forest, 3) Gradient Boosting Tree, 4) Naïve Bayes, 5) Logistic Regression and 6) Voting. The findings revealed that the factors that most affected the decision to study in the bachelor's degree from all techniques of qualification selection were: Selected field of study or course, original course and original qualifications. The technique with the highest accuracy in forecasting was the Voting technique with an accuracy of 73.44%, with a higher accuracy than the Random Forest, Logistic Regression, Decision Tree, Naïve Bayes and Gradient Boosting Tree, whose accuracies were 71.45%, 71.05%, 68.26%, 65.60%, 64.81%, respectively.

คำสำคัญ: เหมืองข้อมูล การคัดเลือกคุณลักษณะ การรับสมัคร การเข้าศึกษาต่อ
Keywords: Data Mining, Feature Selection, Recruitment, Admission

บทนำ

ในปัจจุบันรูปแบบการรับสมัครนักศึกษาใหม่ในแต่ละสถาบันอุดมศึกษาต่างมีรูปแบบที่เป็นนโยบายเชิงรุกและมีการแข่งขันที่มากขึ้น รวมถึงการประชาสัมพันธ์ในหลากหลายช่องทางในเผยแพร่ข่าวสารการรับสมัครนักศึกษา เพื่อให้ผู้สมัครเกิดความสนใจที่จะตัดสินใจเข้าศึกษาต่อในสถาบันอุดมศึกษานั้นๆ ในขณะที่จำนวนนักเรียนในระบบการศึกษาของประเทศไทยมีจำนวนลดลง อันเป็นผลมาจากอัตราการเกิดของประชากรในประเทศไทยลดน้อยลงทุกปี ในปี พ.ศ. 2554 ประเทศไทยยังมีตัวเลขการเกิดสูงเกือบ 8.5 แสนคน ในขณะที่ในปี พ.ศ. 2564 ตัวเลขการเกิดต่ำกว่า 5.5 แสนคน หรือกล่าวอีกนัยหนึ่งว่าตัวเลขการเกิดลดลงมากกว่า 1 ใน 3 โดยใช้เวลาเพียง 10 ปีเท่านั้น (ธีระ, 2565) ทำให้แต่ละสถาบันอุดมศึกษาต้องมีการแข่งขันในการรับสมัครนักศึกษากันมากขึ้นตามไปด้วย การรับสมัครนักศึกษาใหม่ของคณะบริหารธุรกิจและศิลปศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา เชียงใหม่ มีการคัดเลือกอยู่ 4 รูปแบบหลักๆ ด้วยกันคือ 1) การสอบคัดเลือกด้วยวิธียื่นแฟ้มสะสมผลงาน (Portfolio) หรือ TCAS-1 2) การสอบคัดเลือกด้วยวิธียื่นคะแนนและโควตา (Quota) หรือที่เรียกว่ารอบ TCAS-2 3) การสอบคัดเลือกด้วยวิธีแอดมิชชั่น (Admission) หรือ TCAS-3 และ 4) ถ้าสาขาวิชาหรือหลักสูตรใดยังมีจำนวนนักศึกษาที่รับยังไม่ครบตามที่ประกาศรับก็จะมีรอบเพิ่มเติม (Extra) หรือ TCAS-4

สถาบันอุดมศึกษาทั้งในและต่างประเทศมีการแข่งขันรับสมัครนักศึกษากันมากขึ้น ทำให้บางสาขาวิชาหรือหลักสูตรของคณะบริหารธุรกิจฯ มีจำนวนนักศึกษาที่ลดลงและไม่เต็มตามจำนวนที่ประกาศรับรับสมัคร คณะบริหารธุรกิจฯ ได้มีการจัดเก็บข้อมูลผู้สมัครเป็นจำนวนมาก เช่น ข้อมูลสายการเรียนเดิม ข้อมูลวุฒิการศึกษาเดิม ข้อมูลพื้นฐานครอบครัว ข้อมูลเกรดเฉลี่ยเดิม ข้อมูลรูปแบบการสอบคัดเลือก ข้อมูลคณะและสาขาวิชาที่เลือก เป็นต้น ข้อมูลผู้สมัครจำนวนมากเหล่านี้ สามารถนำมาวิเคราะห์เพื่อให้คณะบริหารธุรกิจและศิลปศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา เชียงใหม่ นำไปใช้ประโยชน์ในด้านต่างๆ เช่น การประชาสัมพันธ์สาขาวิชาหรือหลักสูตรและการแนะแนวตามโรงเรียนต่างๆ การนำผลการวิเคราะห์ไปประกอบการวางแผนการรับสมัครนักศึกษาใหม่ในปีการศึกษาต่อไปได้อย่างมีประสิทธิภาพ เพื่อให้สามารถเข้าถึงกลุ่มนักเรียน

เป้าหมายได้มากขึ้น โน้มน้าวให้กลุ่มนักเรียนเข้ามาสมัครเรียนและศึกษาต่อในสาขาวิชาที่ตนเองสนใจในระดับปริญญาตรีของคณะบริหารธุรกิจฯ มากขึ้นตามไปด้วย

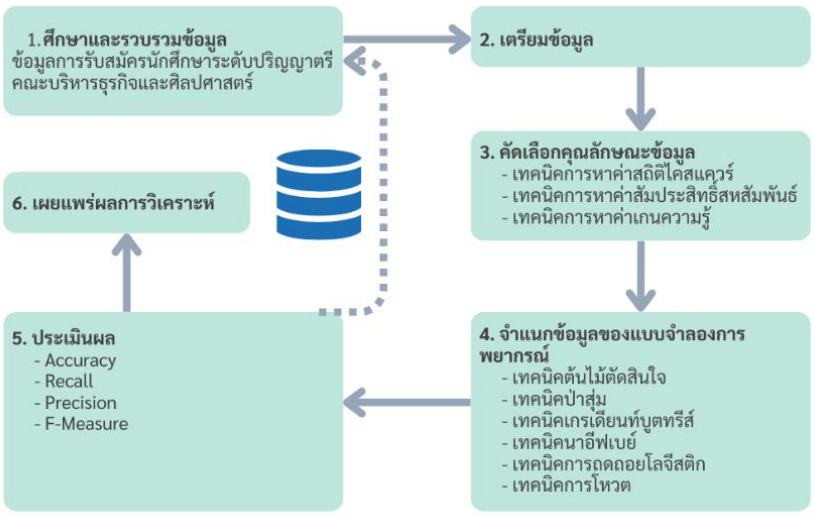
ดังนั้นคณะผู้วิจัยจึงได้ศึกษาปัจจัยที่ส่งผลต่อการตัดสินใจเข้าศึกษาต่อในระดับปริญญาตรี คณะบริหารธุรกิจและศิลปศาสตร์ โดยใช้เทคนิคการคัดเลือกคุณลักษณะและเทคนิคการทำเหมืองข้อมูล มีวัตถุประสงค์เพื่อวิเคราะห์ปัจจัยใดที่ส่งผลต่อการตัดสินใจเข้าศึกษาต่อในระดับปริญญาตรี คณะบริหารธุรกิจและศิลปศาสตร์ และเพื่อเปรียบเทียบประสิทธิภาพการพยากรณ์ด้วยเทคนิคการทำเหมืองข้อมูล โดยนำข้อมูลของผู้สมัครเป็นนักศึกษาในระดับปริญญาตรี คณะบริหารธุรกิจและศิลปศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา เชียงใหม่ ในช่วงปีการศึกษา 2563 – 2565 จำนวน 2,509 ราย มาทำการวิเคราะห์ปัจจัยที่ส่งผลต่อการตัดสินใจเข้าศึกษาต่อในระดับปริญญาตรี คณะบริหารธุรกิจและศิลปศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา เชียงใหม่ โดยใช้เทคนิคการคัดเลือกคุณลักษณะและเทคนิคการทำเหมืองข้อมูล โดยเทคนิคการคัดเลือกคุณลักษณะที่ใช้ประกอบไปด้วย 3 เทคนิค ได้แก่ 1) เทคนิคการหาค่าสถิติไคสแควร์ 2) เทคนิคการหาค่าสัมประสิทธิ์สหสัมพันธ์ และ 3) เทคนิคการหาค่าเกนความรู้ และเทคนิคการทำเหมืองข้อมูลที่ใช้ประกอบไปด้วย 6 เทคนิค ได้แก่ 1) เทคนิคต้นไม้ตัดสินใจ 2) เทคนิคป่าสุ่ม 3) เทคนิคเกรเดียนท์บูตทรีส์ 4) เทคนิคนาอีฟเบย์ 5) เทคนิคการถดถอยโลจีสติก และ 6) เทคนิคการโหวต คาดว่าผลการวิจัยจะเป็นประโยชน์ในด้านของการแนะนำสาขาวิชาหรือหลักสูตร ประกอบการวางแผน กำหนดเงื่อนไขในการรับสมัครงานแนะแนวการศึกษาด้านการประชาสัมพันธ์ตามสถาบันการศึกษาต่างๆ ได้ตรงตามกลุ่มเป้าหมายมากยิ่งขึ้น

วิธีการดำเนินการวิจัย

1. ข้อมูลที่ใช้ในการวิจัยครั้งนี้ได้แก่ ข้อมูลการรับสมัครนักศึกษาในระดับปริญญาตรี คณะบริหารธุรกิจและศิลปศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา เชียงใหม่ ในปี 2563 - 2565 จำนวน 2,509 รายการ

2. เครื่องมือที่ใช้ในการวิจัย ในการวิจัยครั้งนี้คณะผู้วิจัยได้ใช้โปรแกรมประยุกต์ตารางคำนวณ Microsoft Excel ในขั้นตอนการรวบรวมและขั้นตอนของการเตรียมข้อมูล และใช้โปรแกรมประยุกต์ RapidMiner Studio 10.1 (Educational Edition) ในขั้นตอนการคัดเลือกคุณลักษณะ การสร้างแบบจำลองด้วยเทคนิคต่างๆ และขั้นตอนของการประเมินผลประสิทธิภาพของเทคนิคการพยากรณ์

3. ในการศึกษาวิจัยครั้งนี้ คณะผู้วิจัยได้ใช้กระบวนการทำเหมืองข้อมูลตามกรอบแนวคิดของ CRISP-DM (Cross-Industry Standard Process for Data Mining) (Shearer, 2000) ดังรูปที่ 1



รูปที่ 1 กรอบแนวคิดการวิจัย

3.1 ขั้นตอนการศึกษาและรวบรวมข้อมูล (Data Gathering)

คณะผู้วิจัยได้ทำการรวบรวมข้อมูลของผู้สมัครเป็นนักศึกษาในระดับปริญญาตรี คณะบริหารธุรกิจและศิลปศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา เชียงใหม่ ในช่วงปีการศึกษาที่ 2563 – 2565 โดยมีจำนวนข้อมูลผู้สมัครทั้งหมด 2,509 รายการ จากงานส่งเสริมวิชาการและงานทะเบียน มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา เชียงใหม่ โดยคณะผู้วิจัยได้รับไฟล์ข้อมูลมาในรูปแบบไฟล์ csv ข้อมูลแต่ละรายการประกอบด้วย ปีการศึกษาที่รับสมัคร รหัสประจำตัวสอบ เพศ รูปแบบการสอบคัดเลือก ข้อมูลเขตพื้นที่ คณะวิชาที่เลือก หลักสูตรหรือสาขาวิชาที่เลือก สถาบันการศึกษาเดิม สายการเรียนเดิม วุฒิการศึกษาเดิม เกรดเฉลี่ยเดิม ข้อมูลจังหวัดของสถาบันการศึกษาที่สำเร็จการศึกษา ข้อมูลสถานะ การรายงานตัว เป็นต้น แต่เนื่องด้วยตารางข้อมูลสามารถแสดงข้อมูลได้จำนวนจำกัด คณะผู้วิจัยจึงได้แสดงตัวอย่างข้อมูลที่สำคัญบางส่วนในตารางเท่านั้น ดังตารางที่ 1

ตารางที่ 1 ตัวอย่างข้อมูลการรับสมัครนักศึกษาในระดับปริญญาตรี

ปีการศึกษา	เพศ	รูปแบบการสอบคัดเลือก	รหัสหลักสูตรหรือสาขาวิชาที่เลือก	หลักสูตรหรือสาขาวิชาที่เลือก	สถาบันการศึกษาเดิม	สายการเรียนเดิม	วุฒิการศึกษาเดิม	เกรดเฉลี่ยเดิม	จังหวัดของสถาบันการศึกษาเดิม	สถานะการรายงานตัว
2563	ชาย	TCAS1	200	บธ.บ. การตลาด	วิทยาลัยอาชีวศึกษาอุตรดิตถ์	อาชีพ	ปวส.	3.75	อุตรดิตถ์	รายงานตัว
2564	ชาย	TCAS2	201	บธ.บ. การจัดการ	วิทยาลัยอาชีวศึกษาเชียงใหม่	อาชีพ	ปวช.	2.89	เชียงใหม่	ไม่มารายงานตัว
2565	หญิง	TCAS2.1	202	บธ.บ.การท่องเที่ยว	โรงเรียนวัฒโนทัยพายัพ	สามัญ	ม.6	3.50	เชียงใหม่	รายงานตัว

3.2 ขั้นตอนการเตรียมข้อมูล (Data Pre-processing)

จากขั้นตอนการศึกษาและรวบรวมข้อมูล เบื้องต้นคณะผู้วิจัยพบว่าข้อมูลมีความไม่สมบูรณ์และมีค่าที่ขาดหายไป (Missing value) โดยไม่สามารถนำข้อมูลเข้าสู่กระบวนการในขั้นตอนต่อไปได้ คณะผู้วิจัยได้ทำการตัดแถวรายการที่ไม่สมบูรณ์ทิ้ง ได้แก่ ข้อมูลที่ไม่ระบุวุฒิการศึกษาเดิมจำนวน 12 รายการ ข้อมูลที่ไม่ระบุจังหวัดของสถาบันการศึกษาที่สำเร็จการศึกษาจำนวน 143 รายการ และ ข้อมูลที่ไม่ระบุเกรดเฉลี่ยจำนวน 157 รายการ จากเดิมมีจำนวนข้อมูลนักศึกษาทั้งหมด 2,821 รายการ จึงเหลือข้อมูลที่สมบูรณ์จำนวน 2,509 รายการ

จากขั้นตอนการจัดการข้อมูลที่ไม่สมบูรณ์หรือมีค่าที่ขาดหายไปเรียบร้อยแล้ว คณะผู้วิจัยได้ทำการเลือกแอตทริบิวต์ (Feature Selection) เฉพาะที่เกี่ยวข้องและมีความสำคัญ โดยได้ทำการตัดแอตทริบิวต์ ปีการศึกษา สาขาวิชาหรือหลักสูตรที่เลือก สถาบันการศึกษาเดิม และเหลือจำนวนแอตทริบิวต์จำนวน 8 แอตทริบิวต์ ได้แก่ เพศ รูปแบบการรับสมัคร รหัสหลักสูตรหรือสาขาวิชาที่เลือก สายการเรียนเดิม วุฒิการศึกษาเดิม เกรดเฉลี่ยเดิม จังหวัดของสถาบันการศึกษาเดิม และแอตทริบิวต์สำหรับกำหนดเป็นคลาสคำตอบ ได้แก่ สถานะการรายงานตัว

เนื่องจากข้อมูลมีทั้งที่เป็นตัวเลข ตัวอักษร คณะผู้วิจัยจึงต้องทำการปรับเปลี่ยนรูปแบบข้อมูล (Data Transformation) โดยแทนค่าข้อมูลให้อยู่ในรูปแบบที่เหมาะสม เพื่อให้สามารถนำไปวิเคราะห์ได้ตามเทคนิคการทำเหมืองข้อมูล โดยแอตทริบิวต์ที่มีการปรับเปลี่ยนรูปแบบข้อมูลประกอบไปด้วย เพศ สายการเรียนเดิม วุฒิการศึกษาเดิม จังหวัดของสถาบันการศึกษาเดิม สถานะการรายงานตัว ดังตารางที่ 2

ตารางที่ 2 การปรับเปลี่ยนรูปแบบข้อมูล

แอตทริบิวต์	ข้อมูลเดิมก่อนการปรับเปลี่ยนรูปแบบ	ข้อมูลใหม่หลังการปรับเปลี่ยนรูปแบบ
เพศ	ชาย	male
	หญิง	female
สายการเรียนเดิม	สายสามัญ	Ordinary
	สายอาชีพ	Occupation
วุฒิการศึกษาเดิม	ม.6	Senior High School
	ปวช.	Vocational Certificate
	ปวส.	High Vocational Certificate
จังหวัดของสถาบันการศึกษาเดิม	เชียงใหม่	Chiangmai
	...	...
	อุตรดิตย์	Uttaradit
สถานะการรายงานตัว	รายงานตัว	Y
	ไม่มารายงานตัว	N

จากตารางที่ 2 คณะผู้วิจัยได้ทำปรับแก้ในส่วนของชื่อแอตทริบิวต์ใหม่ โดยมีรายละเอียดดังตารางที่ 3

แอตทริบิวต์เพศ เปลี่ยนชื่อเป็น sex (แอตทริบิวต์)

แอตทริบิวต์รูปแบบการสอบคัดเลือก เปลี่ยนชื่อเป็น qualifyExam (แอตทริบิวต์)

แอตทริบิวต์รหัสหลักสูตรหรือสาขาวิชาที่เลือก เปลี่ยนชื่อเป็น courseCode (แอตทริบิวต์)

แอตทริบิวต์สายการเรียนเดิม เปลี่ยนชื่อเป็น oldMajor (แอตทริบิวต์)

แอตทริบิวต์วุฒิการศึกษาเดิม เปลี่ยนชื่อเป็น oldEducational (แอตทริบิวต์)

แอตทริบิวต์เกรดเฉลี่ยเดิม เปลี่ยนชื่อเป็น oldGpa (แอตทริบิวต์)

แอตทริบิวต์จังหวัดของสถาบันการศึกษาเดิม เปลี่ยนชื่อเป็น provinceOfInstitution (แอตทริบิวต์)

แอตทริบิวต์สถานะการรายงานตัว เปลี่ยนชื่อเป็น reportStatus (คลาสคำตอบ)

ตารางที่ 3 ชุดข้อมูลที่ผ่านขั้นตอนการเตรียมข้อมูลสมบูรณ์

sex	qualifyExam	courseCode	oldMajor	oldEducational	oldGpa	province OfInstitution	reportStatus
male	TCAS1	200	Occupation	High Vocational Certificate	3.75	Uttaradit	Y
female	TCAS2	201	Occupation	Vocational Certificate	2.89	Chiangmai	N

หลังจากขั้นตอนการเตรียมข้อมูลเรียบร้อยแล้ว พบว่าข้อมูลมีความไม่สมดุล มีจำนวนกลุ่มนักศึกษาที่มีสถานะการรายงานตัวมากกว่ากลุ่มนักศึกษาที่ไม่มารายงานตัว ซึ่งอาจทำให้ขั้นตอนการวิเคราะห์ไม่มีประสิทธิภาพได้ ดังนั้นคณะผู้วิจัยได้ทำการแก้ปัญหาโดยการนำวิธีการสุ่มเพิ่มตัวอย่างข้อมูลกลุ่มน้อย (SMOTE: Synthetic Minority Over-sampling

Technique) (Chawla *et al*., 2002) เพื่อช่วยในการปรับสมดุลของข้อมูลและเพื่อให้จำนวนข้อมูลของคลาสรอง (Minority class) มีจำนวนเทียบเท่าหรือใกล้เคียงกับจำนวนข้อมูลที่เป็นคลาสหลัก (Majority class) ก่อนนำข้อมูลเข้าสู่กระบวนการในขั้นตอนวิเคราะห์ข้อมูล เทคนิคการสุ่มเพิ่มตัวอย่างข้อมูลกลุ่มน้อย มีหลักการทำงานคล้ายกับอัลกอริทึมเพื่อนบ้านใกล้เคียงที่สุด โดยจะทำการสุ่มเลือกข้อมูลแล้วเก็บในตัวแปร x_i โดยทั่วไปแล้วตัวแปร x_i และตัวแปร $x^{'}$ จะถูกเรียกว่า seed sample จากนั้นจะทำการสุ่มค่าตัวแปร σ ให้มีค่า 0 หรือ 1 เพื่อใช้ในการคำนวณ โดยมีสมการในการสุ่มเพิ่มตัวอย่างข้อมูลกลุ่มน้อย ดังสมการที่ 1

$$x_{new} = x_i + (x^{'} - x_i) \times \sigma \quad (1)$$

โดยที่ x_{new} = ข้อมูลกลุ่มน้อยใหม่ ที่ถูกสุ่มเพิ่มเข้ามาในชุดข้อมูล

ตารางที่ 4 เปรียบเทียบจำนวนข้อมูลที่ผ่านกระบวนการสุ่มเพิ่มตัวอย่างข้อมูลกลุ่มน้อย

กลุ่ม	ข้อมูลก่อนการทำ SMOTE	ข้อมูลหลังการทำ SMOTE
กลุ่มที่รายงานตัว	1,649	1,649
กลุ่มที่ไม่มารายงานตัว	860	1,649
จำนวนทั้งหมด	2,509	3,298

จากตารางที่ 4 เมื่อใช้เทคนิคการสุ่มเพิ่มตัวอย่างข้อมูลกลุ่มน้อย (กลุ่มที่ไม่มารายงานตัว) ทำให้มีจำนวนข้อมูลกลุ่มที่ไม่มารายงานตัวเทียบเท่ากลุ่มที่รายงานตัว 1,649 รายการ จากเดิมที่มีข้อมูลกลุ่มที่ไม่มารายงานตัวจำนวน 860 รายการ

3.3 ขั้นตอนเลือกคุณลักษณะข้อมูล (Feature Selection)

ขั้นตอนการคัดเลือกคุณลักษณะ เป็นขั้นตอนในการวิเคราะห์คุณลักษณะหรือปัจจัยที่ส่งผลต่อการตัดสินใจเข้าศึกษาต่อในระดับปริญญาตรี โดยใช้เทคนิคการหาค่าน้ำหนัก (Filter Approach) จำนวน 3 เทคนิค ได้แก่ 1) เทคนิคการหาค่าสถิติไคสแควร์ 2) เทคนิคการหาค่าสัมประสิทธิ์สหสัมพันธ์ และ 3) เทคนิคการหาค่าเกนความรู้

เทคนิคการหาค่าสถิติไคสแควร์ (Chi-Square Statistic) เป็นเทคนิคการเลือกคุณลักษณะแบบการคัดกรอง (Filter-Based Feature Selection) จะใช้เงื่อนไขทางสถิติในการคัดกรอง โดยค่าสถิติไคสแควร์วัดค่าความสัมพันธ์ระหว่างคุณลักษณะกับคลาสคำตอบ เพื่อจัดลำดับคุณลักษณะตามค่านัยสำคัญทางสถิติ (Jin *et al*., 2006) ดังสมการที่ 2

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad (2)$$

โดยที่ χ^2 = ค่าสถิติ chi-square

O_i = ความถี่ที่ได้จากการสังเกต (Observed Frequency)

E_i = ความถี่ที่คาดหวัง (Expected Frequency) ซึ่งมีค่าเท่ากับจำนวนข้อมูล คูณด้วยสัดส่วนที่คาดหวัง

เทคนิคการหาค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation Based) เป็นเทคนิคในการเลือกคุณสมบัติของคุณลักษณะโดยใช้การพิจารณาบนพื้นฐานความสัมพันธ์ของกลุ่มคุณลักษณะที่ได้จากการประเมินค่าความสามารถในการคาดการณ์ และยังสามารถจัดการกับคุณลักษณะที่ไม่เกี่ยวข้อง โดยจะจัดอันดับกลุ่มย่อยของมิติข้อมูลและทำการคัดเลือกกลุ่มย่อยของมิติข้อมูลที่มีความสัมพันธ์กันสูงกับคลาสและไม่มีความสัมพันธ์กับคลาสอื่นๆ (Hall, 1999) ดังสมการที่ 3

$$M_s = \frac{k\overline{r_{cf}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}} \quad (3)$$

โดยที่ M_s คือค่าค้นหาได้ของมิติข้อมูลกลุ่มย่อย S ซึ่งประกอบไปด้วยมิติข้อมูล k

$\overline{r_{cf}}$ คือค่าเฉลี่ยความสัมพันธ์ของตัวแปรกับคลาส ($f \in s$)

$\overline{r_{ff}}$ คือค่าเฉลี่ยความสัมพันธ์ระหว่างมิติข้อมูล

เทคนิคการหาค่าเกนความรู้ (Information Gain) เป็นเทคนิคการเลือกคุณลักษณะเพื่อประเมินค่าในการแบ่งข้อมูลด้วยการคำนวณค่า Gain สำหรับแต่ละมิติข้อมูล ถ้ามิติข้อมูลใดมีค่า Gain สูงสุด จะถูกเลือกให้เป็นกลุ่มย่อยที่มีอำนาจในการจำแนก (Lee and Lee, 2006) ดังสมการที่ 4

$$\text{Gain} = \text{Entropy}(p) - \sum_{i=1}^{k} \frac{ni}{n} \text{Entropy}(i) \quad (4)$$

โดยที่ $\text{Entropy}(p)$ คือค่า Entropy ของตัว Root

$\sum_{i=1}^{k} \frac{ni}{n} \text{Entropy}(i)$ คือค่า Entropy ในแต่ละโหนดย่อย

คำนวณหาค่า Entropy ได้จากสมการที่ 5

$$\text{Entropy}(p) = \sum_{i=0}^{c-1} p(j|t) \log_2 p(j|t) \quad (5)$$

โดยที่ Σ_i คือผลรวมของความน่าจะเป็นของค่า j ที่เกิดในคลาส t

$p(j|t)$ คือค่าความถี่ที่มีความสัมพันธ์ของกลุ่ม j กับโหนด t

3.4 ขั้นตอนการสร้างแบบจำลอง (Modeling)

ในขั้นตอนนี้เป็นขั้นตอนของการวิเคราะห์ข้อมูลด้วยเทคนิคการทำเหมืองข้อมูล คณะผู้วิจัยได้เลือกใช้วิธีการสร้างแบบจำลองในการวิเคราะห์ข้อมูล โดยใช้เทคนิคการจำแนกข้อมูล (Classification) 6 เทคนิค ได้แก่

เทคนิคต้นไม้ตัดสินใจ (Decision Tree) (Quinlan, 1986) เป็นวิธีการจำแนกประเภทข้อมูลหรือจำแนกกฎ โดยมีลักษณะการทำงานคล้ายกับโครงสร้างต้นไม้ โดยแต่ละโหนด (Node) จะเป็นตัวแทนของคุณลักษณะ (Attribute) ที่ใช้ทดสอบข้อมูลแต่ละกิ่งแสดงผลในการทดสอบ และลีฟโหนด (Leaf node) แสดงถึงกลุ่มคลาสหรือกลุ่มคำตอบ (Class) ที่กำหนดไว้ เทคนิคต้นไม้ตัดสินใจเป็นเทคนิคที่ใช้งานกันอย่างแพร่หลาย เนื่องจากผู้ใช้สามารถทำความเข้าใจผลลัพธ์ได้ง่ายไม่ซับซ้อน โดยเทคนิคต้นไม้ตัดสินใจจะทำการคัดเลือกแอตทริบิวต์ที่มีความสัมพันธ์กับคลาสมากที่สุดขึ้นมาเป็นโหนดบนสุดของต้นไม้ เรียกว่าโหนดราก (Root Node) จากนั้นจะเลือกคุณสมบัติที่มีความสัมพันธ์ถัดไปเรื่อยๆ จากการคำนวณค่าเกนความรู้ โดยเลือกจากคุณสมบัติที่มีค่าเกนความรู้สูงสุด ดังสมการที่ 6

$$IG(parent,child) = Entorpy(parent) - [p(c1) \times Entropy(c1) + p(c2) \times Entropy(c2) + \dots] \quad (6)$$

โดยที่ Entropy(c1) คือ -p(c1) log p(c1)

p(c1) คือค่าความน่าจะเป็นของ c1

c คือปัจจัย (Attribute) แต่ละตัวที่เกี่ยวข้อง

ซึ่งค่า Entropy นี้จะใช้ในการวัดความแตกต่างกันของข้อมูล ถ้าข้อมูลมีความแตกต่างกันน้อย ค่า Entropy จะมีค่าต่ำ แต่ถ้าข้อมูลมีความแตกต่างกันมาก ค่า Entropy จะมีค่าสูง ดังนั้นข้อมูล Entropy ของโหนดลูก (Child Node) สามารถสร้างโมเดลของต้นไม้ตัดสินใจโดยจะคำนวณค่า Information Gain ของแต่ละแอตทริบิวต์เทียบกับคลาสเพื่อหาแอตทริบิวต์ที่มีค่า Information Gain มากที่สุดมาเป็นโหนดราก (Root Node) ของโมเดลต้นไม้ตัดสินใจ

เทคนิคป่าสุ่ม (Random Forest) (Breiman, 2001) เป็นเทคนิคที่พัฒนาต่อจากเทคนิคต้นไม้ตัดสินใจ โดยหลักการทำงานของเทคนิคป่าสุ่มจะทำการสร้างโมเดลจากโมเดลต้นไม้ตัดสินใจ โดยการสร้างต้นไม้หลายๆ ต้น โดยต้นไม้แต่ละต้นจะได้รับข้อมูลสำหรับการเรียนรู้ไม่เหมือนกัน (Subset Data) แต่จะเป็นชุดข้อมูลสำหรับการเรียนรู้เดียวกัน (Data Set) เมื่อได้ผลการทำนายของต้นไม้แต่ละต้นก็จะนำผลการทำนายมาใช้หลักการโหวต จากผลการทำนายต้นไม้ตัดสินใจแต่ละต้นที่ถูกเลือกมากที่สุด ก็จะเป็นคำตอบของคลาสที่กำลังพิจารณาอยู่

เทคนิคเกรเดียนท์บูตทรีส์ (Gradient Boosting Tree) (Natekin and Knoll, 2013) เป็นเทคนิคที่มีพื้นฐานมาจากเทคนิคต้นไม้ตัดสินใจ โดยเป็นการปรับปรุงประสิทธิภาพของแบบจำลองต้นไม้ตัดสินใจให้มีประสิทธิภาพมากขึ้น โดยการสุ่มสร้างต้นไม้ตัดสินใจหลายสิบต้นจนถึงหลายร้อยต้น และทำการประเมินผลต้นไม้แต่ละต้นจนกว่าจะได้ต้นไม้ตัดสินใจที่มีประสิทธิภาพมากที่สุด

เทคนิคนาอีฟเบย์ (Naïve Bayes) (Rish, 2001) เป็นเทคนิคที่ใช้ความน่าจะเป็น (Probability) ในการจำแนกข้อมูลตามกฎของเบย์ (Bayes' Theorem) เพื่อหาว่าสมมติฐานใดน่าจะถูกต้องที่สุด โดยใช้ความรู้ก่อนหน้า (Prior Knowledge) ได้แก่ ความน่าจะเป็นก่อนหน้าสำหรับสมมติฐานหนึ่งๆ ร่วมกับข้อมูล เช่น ความน่าจะเป็นที่สังเกตได้สำหรับสมมติฐานหนึ่งๆ เพื่อหาสมมติฐานที่ดีที่สุด เป็นเทคนิคที่ได้รับความนิยมอย่างแพร่หลายในปัจจุบัน ดังสมการที่ 7

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{7}$$

โดยที่ $P(A|B)$ คือค่าความน่าจะเป็นที่เกิดเหตุการณ์ B ขึ้นก่อนและจะมีเหตุการณ์ A ตามมา (Conditional Probability)
$P(A \cap B)$ คือค่าความน่าจะเป็นที่เหตุการณ์ A และเหตุการณ์ B จะเกิดร่วมกัน (Joint Probability)
$P(B)$ คือค่าความน่าจะเป็นที่เหตุการณ์ B จะเกิดขึ้น

เทคนิคการถดถอยโลจีสติก (Logistic Regression) (Peng *et al*., 2002) เป็นเทคนิคการวิเคราะห์ตัวแปรเชิงพหุที่มีวัตถุประสงค์เพื่อประมาณค่าหรือทำนายเหตุการณ์ที่สนใจว่าจะเกิดหรือไม่เกิดเหตุการณ์นั้นภายใต้ปัจจัยต่างๆ แบบจำลองโลจีสติกประกอบไปด้วยตัวแปรตาม (Dependent Variable) ที่ต้องเป็นตัวแปรตามแบบทวินาม (Dichotomous Variable) คือมีค่าได้สองค่า ตัวอย่างเช่น "ฝนตก" กับ "ไม่ตก" "รายงานตัว" กับ "ไม่รายงานตัว" เป็นต้น และมีตัวแปรอิสระ (Independent Variable) หรือเรียกว่าตัวแปรทำนาย (Predictor) ที่อาจมีตัวเดียวหรือหลายตัวที่เป็นไปได้ทั้งตัวแปรเชิงกลุ่ม (Categorical Variable) หรือตัวแปรแบบต่อเนื่อง (Continuous Variable) เทคนิคการถดถอยโลจีสติกเป็นการประมาณค่าความน่าจะเป็นของการเกิดเหตุการณ์ มีตัวแบบมาจากฟังก์ชันโลจีสติก ถ้าหากพิจารณาตามจำนวนตัวแปรอิสระ ถ้ามีตัวแปรอิสระ 1 ตัว เรียกว่า การวิเคราะห์การถดถอยโลจีสติกอย่างง่าย (Simple Logistic Regression) แต่ถ้ามีตัวแปรอิสระ 2 ตัวขึ้นไป เรียกว่า การวิเคราะห์การถดถอยโลจีสติกเชิงพหุ (Multiple Logistic Regression)

เนื่องจากตัวแปรตาม เป็นตัวแปรเชิงคุณภาพที่มีค่าได้ 2 ค่า คือไม่เกิดเหตุการณ์ (Y = 0) หรือเกิดเหตุการณ์ (Y = 1) จึงทำให้ความสัมพันธ์ระหว่างตัวแปรตาม (Y) กับตัวแปรอิสระ (X) ไม่อยู่ในรูปเชิงเส้น แต่จะอยู่ในรูปดังสมการที่ 8 และสมการที่ 9 (กาญจน์เขจร, 2561)

$$\text{Prob (event)} = \frac{e^{Z}}{1 + e^{Z}} \quad (8)$$

หรือ

$$\text{Prob (event)} = \frac{1}{1 + e^{-Z}} \quad (9)$$

เมื่อ Z คือ Linear Combination ที่อยู่ในรูปดังสมการที่ 10

$$Z = \beta_0 + \beta_1 x_1 + \beta_1 x_1 + \ \ldots + \beta_n x_n \quad (10)$$

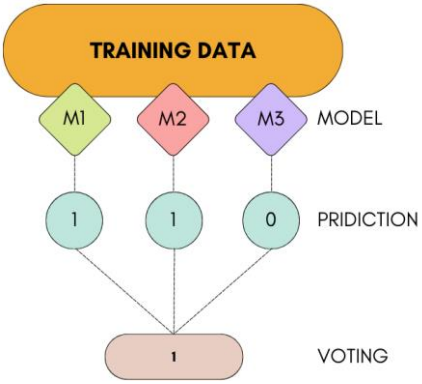
โดยที่ β_0 คือ ค่าคงที่ (เมื่อไม่มีอิทธิพลจากตัวแปรอิสระใด)

β_1 คือ ค่าสัมประสิทธิ์ของตัวแปรอิสระ (ประมาณได้จากข้อมูลสังเกต)

x คือ ตัวแปรอิสระ (ตัวแปรทำนาย)

e คือ ลอการิทึมธรรมชาติ (มีค่าประมาณ 2.71828...)

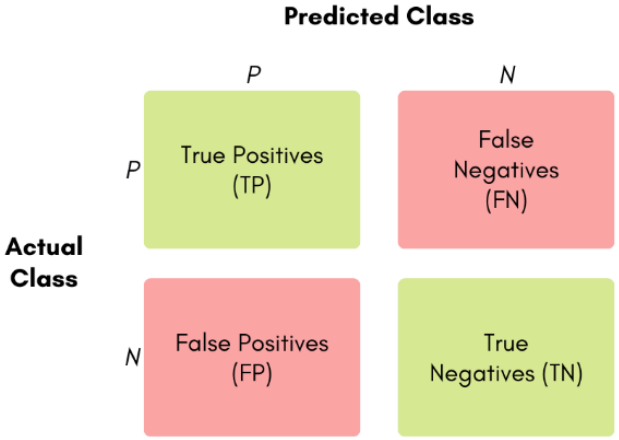
เทคนิคการโหวต (Voting) (Bauer and Kohavi, 1999) เป็นเทคนิคที่จัดอยู่ในการสร้างแบบจำลองการเรียนรู้แบบรวมกลุ่ม (Ensemble Learning-based Model) เป็นการสร้างแบบจำลองการเรียนรู้โดยการใช้โมเดลหรือตัวจำแนก (Classifier) มากกว่าหนึ่งตัวในการเรียนรู้ เทคนิคการโหวตเป็นการนำข้อมูลการเรียนรู้ชุดเดียวกันให้โมเดลแต่ละโมเดลทำการเรียนรู้ เมื่อได้ผลการทำนายของแต่ละโมเดลเรียบร้อยแล้ว จะนำผลการทำนายมาใช้หลักการโหวตว่าคลาสคำตอบใดที่แต่ละโมเดลทำนายผลมากที่สุด ก็จะนำคลาสคำตอบนั้นเป็นผลการทำนายของคลาสที่กำลังพิจารณาอยู่ ดังรูปที่ 2 โดยในงานวิจัยนี้คณะผู้วิจัยได้ใช้โมเดลในการร่วมกันโหวตจำนวน 5 โมเดลด้วยกันคือ เทคนิคต้นไม้ตัดสินใจ เทคนิคป่าสุ่ม เทคนิคเกรเดียนท์บูตทรีส์ เทคนิคนาอีฟเบย์ และเทคนิคการถดถอยโลจีสติก



รูปที่ 2 การทำงานของเทคนิคการโหวต (Voting)

3.5 ขั้นตอนการประเมินผล (Evaluation)

ในขั้นตอนนี้คณะผู้วิจัยได้ทำการประเมินผลการใช้เทคนิคการวิเคราะห์ข้อมูลทั้งหมด 6 เทคนิค และทำการเปรียบเทียบประสิทธิภาพของแต่ละเทคนิค โดยคณะผู้วิจัยได้เลือกใช้วิธีการแบ่งข้อมูลของการทดลองโดยใช้วิธีการแบ่งข้อมูล 70:30 โดยแบ่งข้อมูลสำหรับการเรียนรู้ (Training Data) 70% และข้อมูลสำหรับการทดสอบ (Testing Data) 30% (Vrigazova, 2021) โดยใช้วิธีการประเมินผลประสิทธิภาพของโมเดลด้วยตาราง Confusion Matrix (Sripaoray and Sinsomboonthong, 2017) โดยมีรายละเอียดดังรูปที่ 3



รูปที่ 3 ตาราง Confusion Matrix

โดยที่ TP (True Positive) คือ ผลการทำนายตรงกับสิ่งที่เกิดขึ้นจริง ในกรณีที่ทำนายว่าจริงและสิ่งนั้นเกิดขึ้นจริง

FP (False Positive) คือ ผลการทำนายตรงกับสิ่งที่เกิดขึ้น ในกรณีที่ทำนายว่าจริง และสิ่งที่เกิดขึ้นไม่จริง

TN (True Negative) คือ ผลการทำนายตรงกับสิ่งที่เกิดขึ้น ในกรณีที่ทำนายว่าไม่จริงและสิ่งที่เกิดขึ้นไม่จริง

FN (False Negative) คือ ผลการทำนายไม่ตรงกับสิ่งที่เกิดขึ้นจริง ในกรณีที่ทำนายว่าไม่จริง แต่สิ่งนั้นเกิดขึ้นจริง

โดยบริบทของงานวิจัยนี้ TP คือ กรณีที่นำนายว่านักศึกษามารายงานตัว และนักศึกษามารายงานตัวจริง

FP คือ กรณีที่ทำนายว่านักศึกษามารายงานตัว แต่นักศึกษาไม่มารายงานตัว

TN คือ กรณีที่ทำนายว่านักศึกษาไม่มารายงานตัว และนักศึกษาไม่มารายงานตัวจริง

FN คือ กรณีที่ทำนายว่านักศึกษาไม่มารายงานตัว แต่นักศึกษามารายงานตัว

จากตาราง Confusion Matrix สามารถนำมาคำนวณเพื่อหาค่าการวัดประสิทธิภาพของโมเดลดังนี้

ค่าความถูกต้อง (Accuracy) วิธีวัดความถูกต้องจากจำนวนคำตอบที่ถูกต้อง เทียบกับจำนวนคำตอบทั้งหมดที่นำไปให้โมเดลทำการตอบ โดยใช้สูตรคำนวณ Accuracy = (TP + TN) / (TP + FP + TN + FN)

ค่าความถูกต้อง (Recall) วิธีวัดค่าสัดส่วนของข้อมูลที่ตรงตามความต้องการที่ถูกค้นคืนกับข้อมูลที่ต้องการทั้งหมด โดยใช้สูตรคำนวณ Recall = TP / (TP + FN)

ค่าความแม่นยำ (Precision) วิธีวัดค่าจำนวนที่ทายถูก จากข้อมูลที่ทำนายว่าเป็นคลาสที่พิจารณาอยู่ โดยใช้สูตรคำนวณ Precision = TP / (TP + FP)

ค่าความถ่วงดุล (F-Measure) วิธีการหาค่าน้ำหนักหรือค่าเฉลี่ยของ ค่าความแม่นยำ (Precision) และ ค่าเรียกกลับ (Recall) หรือเป็นการหาประสิทธิภาพโดยรวมของคลาสคำตอบ โดยใช้สูตรคำนวณ F-Measure = 2 x (Precision x recall) / (Precision + Recall)

3.6 ขั้นตอนการเผยแพร่ผลการวิเคราะห์ (Publication)

หลังจากขั้นตอนการประเมินผลแบบจำลองการพยากรณ์การจำแนกข้อมูลของชุดข้อมูลการสมัครเข้าศึกษาต่อในระดับปริญญาตรี คณะบริหารธุรกิจและศิลปศาสตร์เรียบร้อยแล้ว ผู้วิจัยจะเผยแพร่ผลการวิเคราะห์ เพื่อให้หน่วยงานภายในคณะฯ หน่วยงานรับสมัครนักศึกษา นำผลลัพธ์ไปใช้ประโยชน์ด้านของการแนะนำสาขาวิชาหรือหลักสูตร ประกอบการวางแผนหรือเป็นแนวทางในการปรับปรุงและพัฒนาในส่วนของการรับสมัครนักศึกษา ของคณะบริหารธุรกิจและศิลปศาสตร์ต่อไป

ผลการวิจัยและวิจารณ์ผล

1. ผลการศึกษาคุณลักษณะที่ส่งผลต่อการเข้าศึกษาต่อในระดับปริญญาตรี คณะบริหารธุรกิจและศิลปศาสตร์

ผลการศึกษาคุณลักษณะที่ส่งส่งผลต่อการเข้าศึกษาต่อในระดับปริญญาตรี คณะบริหารธุรกิจและศิลปศาสตร์ โดยใช้เทคนิคการคำนวณหาค่าน้ำหนักของคุณลักษณะ ได้แก่ 1) เทคนิคการหาค่าสถิติไคสแควร์ 2) เทคนิคการหาค่าสัมประสิทธิ์สหสัมพันธ์ และ 3) เทคนิคการหาค่าเกนความรู้ สามารถแสดงผลค่าน้ำหนักสูงสุดที่คำนวณได้ 5 ลำดับแรกของแต่ละเทคนิค โดยเรียงตามลำดับหมายเลขในวงเล็บ และนับจำนวนความถี่ที่ได้ 5 ลำดับแรกของแต่ละคุณลักษณะจากการใช้เทคนิคการคัดเลือกคุณลักษณะ ดังตารางที่ 5

ตารางที่ 5 คุณลักษณะ 5 อันดับแรกที่มีความสำคัญมากที่สุดของแต่ละเทคนิค

คุณลักษณะ	ค่าสถิติไคสแควร์	ค่าสัมประสิทธิ์สหสัมพันธ์	ค่าเกนความรู้	ความถี่
เพศ				0
รูปแบบการสอบคัดเลือก				0
รหัสหลักสูตรหรือสาขาวิชาที่เลือก	186.744 (1)	0.067 (4)	0.060 (1)	3
สายการเรียนเดิม	78.270 (3)	0.177 (1)	0.023 (3)	3
วุฒิการศึกษาเดิม	80.260 (2)	0.175 (2)	0.023 (2)	3
เกรดเฉลี่ยเดิม	51.344 (5)	0.104 (3)	0.018 (5)	3
จังหวัดของสถาบันการศึกษาเดิม	67.009 (4)	0.054 (5)	0.020 (4)	3

ผลการศึกษาพบว่าคุณลักษณะ 5 อันดับแรกจากผลการคำนวณหาค่าน้ำหนักสูงสุดทั้ง 3 เทคนิคให้ผลลัพธ์ที่ใกล้เคียงกัน คุณลักษณะที่มีความถี่เท่ากับ 3 ได้แก่ รหัสหลักสูตรหรือสาขาวิชาที่เลือก สายการเรียนเดิม วุฒิการศึกษาเดิม เกรดเฉลี่ยเดิม และจังหวัดของสถาบันการศึกษาเดิม คุณลักษณะที่ไม่มีความถี่ 5 อันดับแรกจากทั้ง 3 เทคนิค คือ เพศ และรูปแบบการสอบคัดเลือก

จากตารางที่ 5 ความถี่แยกตามลำดับของแต่ละเทคนิค โดยคณะผู้วิจัยได้วัดค่าความถี่จากค่าน้ำหนักสูงสุดของแต่ละเทคนิคการคัดเลือกคุณลักษณะ จึงสรุปได้ว่าปัจจัยที่มีความสำคัญ 3 อันดับแรก ได้แก่ อันดับที่ 1 รหัสหลักสูตรหรือสาขาวิชาที่เลือก โดยมีความถี่มาเป็นลำดับที่ 1 จากเทคนิคการหาค่าสถิติไคสแควร์และค่าเกนความรู้ อันดับที่ 2 สายการเรียนเดิม โดยมีความถี่มาเป็นลำดับที่ 1 จากเทคนิคการหาค่าสัมประสิทธิ์สหสัมพันธ์ และอันดับที่ 3 วุฒิการศึกษาเดิม โดยมีความถี่มาเป็นลำดับที่ 2 จากเทคนิคการหาค่าสถิติไคสแควร์ เทคนิคการหาค่าเกนความรู้ และ เทคนิคการหาค่าสัมประสิทธิ์สหสัมพันธ์ตามลำดับ ซึ่งผลการวิจัยเป็นไปในทิศทางเดียวกับ รัชฎา และจรัญ (2563) และ อัศวิน และวรภัทร (2564) พบว่าปัจจัยที่ส่งผลต่อการตัดสินใจเข้าศึกษาต่อหรือเลือกสาขาวิชาในระดับปริญญาตรี ได้แก่ สาขาวิชาที่เลือก และสายการเรียนเดิม

เนื่องด้วยคุณลักษณะมีจำนวนน้อยและอาจทำให้การสร้างแบบจำลองพยากรณ์มีประสิทธิภาพต่ำ คณะผู้วิจัยจึงไม่ได้ตัดคุณลักษณะ เพศ และ รูปแบบการสอบคัดเลือกทิ้งแต่อย่างใด ก่อนนำไปออกแบบและสร้างแบบจำลองพยากรณ์

2. ผลการเปรียบเทียบประสิทธิภาพการจำแนกข้อมูล

สรุปผลเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจ เทคนิคนาอีฟเบย์ เทคนิคป่าสุ่ม เทคนิคการถดถอยโลจีสติก เทคนิคเกรเดียนท์บูตทรีส์ และเทคนิคการโหวต โดยคณะผู้วิจัยได้ทำการประเมินผลประสิทธิภาพของโมเดล โดยได้ทำการเปรียบเทียบกับชุดข้อมูลก่อนการปรับสมดุลข้อมูลและหลังการปรับสมดุลข้อมูลด้วยวิธีการสุ่มเพิ่มตัวอย่างข้อมูลกลุ่มน้อย (SMOTE) ดังตารางที่ 6

ตารางที่ 6 เปรียบเทียบการประเมินผลประสิทธิภาพแต่ละโมเดลก่อนการปรับสมดุลข้อมูล

Model	Accuracy	Recall	Precision	F-Measure
ต้นไม้ตัดสินใจ	68.26%	57.58%	64.46%	60.83%
ป่าสุ่ม	71.45%	64.36%	68.28%	66.26%
เกรเดียนท์บูตทรีส์	64.81%	65.81%	64.26%	65.03%
นาอีฟเบย์	65.60%	96.60%	66.30%	78.63%
การถดถอยโลจีสติก	71.05%	63.13%	68.01%	65.48%
การโหวต	**73.44%**	64.21%	73.57%	68.57%

จากตารางที่ 6 โมเดลที่มีประสิทธิภาพและค่าความถูกต้อง (Accuracy) สูงสุดก่อนการปรับสมดุลข้อมูลคือการใช้เทคนิคการโหวต โดยมีค่าความถูกต้องที่ร้อยละ 73.44% ซึ่งผลการวิจัยเป็นไปในทิศทางเดียวกับ จิราภาและคณะ (2558) และ ปพิชญา (2561) ที่พบว่าการใช้เทคนิคการเรียนรู้แบบรวมกลุ่ม (Ensemble Learning) ด้วยวิธีการโหวต จะให้ค่าความถูกต้องสูงสุดกว่าการใช้แบบจำลองการเรียนรู้แบบเชิงเดี่ยว ส่วนเทคนิคป่าสุ่ม เทคนิคการถดถอยโลจีสติก เทคนิคต้นไม้ตัดสินใจ เทคนิคนาอีฟเบย์ และเทคนิคเกรเดียนท์บูตทรีส์ มีค่าความถูกต้องที่ร้อยละ 71.45% 71.05% 68.26% 65.60% 64.81% ตามลำดับ

ตารางที่ 7 เปรียบเทียบการประเมินผลประสิทธิภาพแต่ละโมเดลหลังการปรับสมดุลข้อมูล

Model	Accuracy	Recall	Precision	F-Measure
ต้นไม้ตัดสินใจ	67.47%	67.47%	67.81%	67.64%
ป่าสุ่ม	**70.81%**	70.81%	71.00%	70.90%
เกรเดียนท์บูตทรีส์	66.77%	66.77%	67.95%	67.35%
นาอีฟเบย์	63.03%	63.03%	63.08%	63.05%
การถดถอยโลจีสติก	65.66%	65.66%	65.66%	65.66%
การโหวต	69.09%	63.03%	69.32%	66.03%

จากตารางที่ 7 โมเดลที่มีประสิทธิภาพและค่าความถูกต้อง (Accuracy) สูงสุดหลังการปรับสมดุลข้อมูลคือการใช้เทคนิคป่าสุ่ม โดยมีค่าความถูกต้องที่ร้อยละ 70.81% และเทคนิคการโหวต เทคนิคต้นไม้ตัดสินใจ เทคนิคเทคนิคเกรเดียนท์บูตทรีส์ เทคนิคการถดถอยโลจีสติก และเทคนิคนาอีฟเบย์ มีค่าความถูกต้องที่ร้อยละ 69.09% 67.47% 66.77% 65.66% 63.03% ตามลำดับ

ตารางที่ 8 เปรียบเทียบค่าความถูกต้องก่อนการปรับสมดุลข้อมูลและหลังการปรับสมดุลข้อมูล

Model	Accuracy	
	Original Data Set	SMOTE Data Set
ต้นไม้ตัดสินใจ	**68.26%**	67.47%
ป่าสุ่ม	**71.45%**	70.81%
เกรเดียนท์บูตทรีส์	64.81%	**66.77%**
นาอีฟเบย์	**65.60%**	63.03%
การถดถอยโลจีสติก	**71.05%**	65.66%
การโหวต	**73.44%**	69.09%

จากตารางที่ 8 จะเห็นได้ว่าข้อมูลหลังการปรับสมดุลข้อมูลโดยวิธีการสุ่มเพิ่มตัวอย่างข้อมูลกลุ่มน้อยมีประสิทธิภาพค่าความถูกต้องน้อยกว่าข้อมูลก่อนการปรับสมดุลข้อมูล มีเพียงเทคนิคเกรเดียนท์บูตทรีส์ที่มีค่าความถูกต้องมากกว่าที่ร้อยละ 66.77% ต่อ 64.81% หลังการปรับสมดุลข้อมูล จึงสรุปได้ว่าการปรับสมดุลข้อมูลจากข้อมูลการรับสมัครนักศึกษาในระดับปริญญาตรี คณะบริหารธุรกิจและศิลปศาสตร์ โดยวิธีการสุ่มเพิ่มตัวอย่างข้อมูลกลุ่มน้อยไม่มีผลทำให้โมเดลมีประสิทธิภาพเพิ่มขึ้นอย่างมีนัยสำคัญ

สรุปผลการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาปัจจัยที่ส่งผลต่อการตัดสินใจเข้าศึกษาต่อในระดับปริญญาตรี คณะบริหารธุรกิจและศิลปศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา เชียงใหม่ และเปรียบเทียบประสิทธิภาพของตัวแบบพยากรณ์โดยใช้เทคนิคจำนวน 6 เทคนิค คือ 1) เทคนิคต้นไม้ตัดสินใจ 2) เทคนิคป่าสุ่ม 3) เทคนิคเกรเดียนท์บูตทรีส์ 4) เทคนิคนาอีฟเบย์ 5) เทคนิคการถดถอยโลจีสติก และ 6) เทคนิคการโหวต โดยการกระบวนการวิจัยเป็นไปตามขั้นตอนของ CRISP-DM ผลการศึกษาพบว่า

1. คุณลักษณะที่ส่งผลต่อการตัดสินใจเข้าศึกษาต่อในระดับปริญญาตรี

จากการใช้เทคนิคการคัดเลือกคุณลักษณะในการคำนวณหาค่าน้ำหนักจำนวน 3 เทคนิค ซึ่งให้ผลลัพธ์ที่ใกล้เคียงกัน พบว่าคุณลักษณะที่มีความถี่ระดับสูงสุดที่ความถี่ 3 มีจำนวน 5 คุณลักษณะ ได้แก่ รหัสหลักสูตรหรือสาขาวิชาที่เลือก สายการเรียนเดิม วุฒิการศึกษาเดิม เกรดเฉลี่ยเดิม และจังหวัดของสถาบันการศึกษาเดิม คุณลักษณะที่ไม่มีความถี่ 5 อันดับแรกจากทั้ง 3 เทคนิค คือ เพศ และ รูปแบบการสอบคัดเลือก ส่วนปัจจัยที่มีความสำคัญ 3 อันดับแรก โดยวัดจากค่าน้ำหนักสูงสุดของแต่ละเทคนิคการคัดเลือกคุณลักษณะ ได้แก่ อันดับที่ 1 รหัสหลักสูตรหรือสาขาวิชาที่เลือก โดยมีความถี่มาเป็นลำดับที่ 1 จากเทคนิคการหาค่าสถิติไคสแควร์ และเทคนิคการหาค่าเกนความรู้ อันดับที่ 2 สายการเรียนเดิม โดยมีความถี่มาเป็นลำดับที่ 1 จากเทคนิค เทคนิคการหาค่าสัมประสิทธิ์สหสัมพันธ์ และอันดับที่ 3 วุฒิการศึกษาเดิม โดยมีความถี่มาเป็นลำดับที่ 2 จากเทคนิคการหาค่าสถิติไคสแควร์ เทคนิคการหาค่าเกนความรู้ และเทคนิคการหาค่าสถิติไคสแควร์

2. ผลการคำนวณประสิทธิภาพการทำนายของวิธีการจำแนกข้อมูล

ประสิทธิภาพการจำแนกข้อมูล ด้วยเทคนิคต้นไม้ตัดสินใจ เทคนิคป่าสุ่ม เทคนิคเกรเดียนท์บูตทรีส์ เทคนิคนาอีฟเบย์ เทคนิคการถดถอยโลจีสติก และเทคนิคการโหวต โดยคำนวณค่าความแม่นยำ ค่าเรียกกลับ ค่าความถูกต้อง ค่าประสิทธิภาพโดยรวม ผลการเปรียบเทียบประสิทธิภาพของแบบจำลองการพยากรณ์พบว่า เทคนิคที่มีประสิทธิภาพและค่าความถูกต้องสูงสุดก่อนการปรับสมดุลข้อมูลด้วยเทคนิคการสุ่มเพิ่มตัวอย่างข้อมูลกลุ่มน้อย คือการใช้เทคนิคการโหวต โดยมีค่าความถูกต้องที่ร้อยละ 73.44% ถือว่าเป็นค่าความถูกต้องที่ยอมรับได้บนชุดข้อมูลนี้ ส่วนเทคนิคป่าสุ่ม เทคนิคการถดถอยโลจีสติก เทคนิคต้นไม้ตัดสินใจ เทคนิคนาอีฟเบย์ และเทคนิคเทคนิคเกรเดียนท์บูตทรีส์ มีค่าความถูกต้องที่ร้อยละ 71.45% 71.05% 68.26% 65.60% 64.81% ตามลำดับ และเทคนิคที่มีประสิทธิภาพและค่าความถูกต้องสูงสุดหลังการปรับสมดุลข้อมูล ด้วยเทคนิคการสุ่มเพิ่มตัวอย่างข้อมูลกลุ่มน้อย (SMOTE) คือการใช้เทคนิคป่าสุ่ม โดยมีค่าความถูกต้องที่ร้อยละ 70.81% และเทคนิคโหวต เทคนิคต้นไม้ตัดสินใจ เทคนิคเกรเดียนท์บูตทรีส์ เทคนิคการถดถอยโลจีสติก และเทคนิคนาอีฟเบย์ มีค่าความถูกต้องที่ร้อยละ 69.09% 67.47% 66.77% 65.66% 63.03% ตามลำดับ ข้อมูลหลังการปรับสมดุลข้อมูลโดยวิธีการสุ่มเพิ่มตัวอย่างข้อมูลกลุ่มน้อยมีประสิทธิภาพค่าความถูกต้องน้อยกว่าข้อมูลก่อนการปรับสมดุลข้อมูล มีเพียงเทคนิคเกรเดียนท์บูตทรีส์ที่มีค่าความถูกต้องมากกว่าที่ร้อยละ 66.77% ต่อ 64.81% จึงสรุปได้ว่าการปรับสมดุลข้อมูลจากข้อมูลการรับสมัครนักศึกษาในระดับปริญญาตรี โดยวิธีการสุ่มเพิ่มตัวอย่างข้อมูลกลุ่มน้อย ไม่มีผลทำให้โมเดลมีประสิทธิภาพเพิ่มขึ้น แต่จะทำให้โมเดลส่วนใหญ่ลดประสิทธิภาพลงอย่างมีนัยสำคัญ โดยคณะผู้วิจัยได้ตั้งข้อสังเกตว่า 1) เมื่อมีการสุ่มเพิ่มตัวอย่างข้อมูลกลุ่มน้อย ทำให้โมเดลให้ค่าน้ำหนักในการเรียนรู้ไปที่คลาสรองเป็นหลักมากกว่า (กลุ่มที่ไม่มารายงานตัว) จึงทำให้โมเดล

เกิดความลำเอียง (bias) เนื่องจากโมเดลจะทำนายคลาสรองด้วยความแม่นยำสูงขึ้น แต่ค่าความถูกต้องโดยรวมจะลดลง ส่วนชุดข้อมูลที่ไม่มีการปรับสมดุลข้อมูล โมเดลจะเรียนรู้การทำนายของคลาสหลักที่มากกว่า ทำให้การพยากรณ์มีความถูกต้องโดยรวมสูงกว่า 2) ข้อมูลอาจมีความยากในการจำแนกข้อมูล ทำให้การพยากรณ์ชุดข้อมูลที่ไม่มีการปรับสมดุลข้อมูลและชุดข้อมูลที่มีการปรับสมดุลข้อมูลมีประสิทธิภาพต่ำ

จากผลการวิจัย ปัจจัยที่ส่งผลต่อการตัดสินใจเข้าศึกษาต่อในระดับปริญญาตรี คณะบริหารธุรกิจและศิลปศาสตร์ โดยใช้เทคนิคการคัดเลือกคุณลักษณะและเทคนิคการทำเหมืองข้อมูล ทำให้สามารถนำผลการวิจัยด้านของปัจจัยที่ส่งผลต่อการตัดสินใจเข้าศึกษาต่อในระดับปริญญาตรีมากที่สุด ได้แก่ หลักสูตรหรือสาขาวิชาที่เลือก สายการเรียนเดิม วุฒิการศึกษาเดิม ไปใช้ประโยชน์ในด้านของการแนะนำสาขาวิชาหรือหลักสูตร ประกอบการวางแผน กำหนดเงื่อนไขในการรับสมัครงานแนะแนวการศึกษาด้านการประชาสัมพันธ์ตามสถาบันการศึกษาต่างๆ ได้ตรงตามกลุ่มเป้าหมายมากยิ่งขึ้น รวมถึงแนวทางในการปรับปรุงพัฒนาในส่วนของงานรับสมัครนักศึกษาของคณะบริหารธุรกิจและศิลปศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา เชียงใหม่ ให้มีประสิทธิภาพมากขึ้น ส่วนเทคนิคการพยากรณ์ด้วยเทคนิคการโหวต เป็นเทคนิคที่เหมาะสมมากที่สุดสำหรับนำไปใช้งานในการพยากรณ์ต่อไป เนื่องจากมีประสิทธิภาพสูงสุดจากทุกเทคนิค

ข้อเสนอแนะการวิจัยครั้งต่อไป

การใช้เทคนิคเหมืองข้อมูลเป็นการค้นหาองค์ความรู้ (Knowledge) จากข้อมูลที่มีอยู่มาทำการวิเคราะห์ปัจจัยที่ส่งผลต่อการตัดสินใจเข้าศึกษาต่อในระดับปริญญาตรี คณะบริหารธุรกิจและศิลปศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา เชียงใหม่ ซึ่งในการวิจัยครั้งต่อไป คณะผู้วิจัยมีความเห็นว่า

1. ควรเพิ่มปัจจัยอื่นๆ ที่อาจจะมีผลต่อการเข้าศึกษาต่อในระดับปริญญาตรี เช่น ปัจจัยด้านผู้ปกครอง อาชีพผู้ปกครอง รายได้ผู้ปกครอง ปัจจัยด้านข้อมูลส่วนตัวนักศึกษา เช่น งานอดิเรก ความสามารถ อาชีพในอนาคต เป็นต้น

2. ควรมีการใช้เทคนิคการพยากรณ์อื่นๆ เข้ามาเปรียบเทียบ เช่น เทคนิคการเรียนรู้แบบรวมกลุ่ม (Ensemble Model) ประเภทต่างๆ เพื่อเพิ่มประสิทธิภาพให้กับตัวแบบพยากรณ์ รวมถึงมีการปรับปรุงพารามิเตอร์ของแต่ละเทคนิคการพยากรณ์ให้ประสิทธิภาพมากขึ้น

3. ควรมีการใช้เทคนิคการปรับสมดุลข้อมูลที่หลากหลาย เพื่อเพิ่มประสิทธิภาพของตัวแบบพยากรณ์ เช่น วิธีการสุ่มเพิ่ม (Oversampling) วิธีการผสมผสาน (Hybrid Method) เป็นต้น

4. ควรมีการทดสอบเทคนิคการคัดเลือกคุณลักษณะและเทคนิคการพยากรณ์หลายๆ ครั้ง เพื่อให้ได้ประสิทธิภาพมากขึ้นและเพื่อไม่ให้เกิดปัญหา overfitting

เอกสารอ้างอิง

กาญจน์เขจร ชูชีพ. (2561). การถดถอยโลจีสติก (Logistic Regression): Remote Sensing Technical Note. คณะวนศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ (5): 1 - 10.

จิราภา เลาหะวรนันท์, รชต ลิ้มสุทธิวันภูมิ และบัณฑิต ฐานะโสภณ. (2558). การใช้เทคนิคการทำเหมืองข้อมูลในการจำแนกและคัดเลือกแขนงวิชาสำหรับนักศึกษาคณะเทคโนโลยีสารสนเทศ. วารสารเทคโนโลยีสารสนเทศลาดกระบัง 4(2): 1 - 9.

ธีระ สินเดชารักษ์. (2565). ชำแหละผลกระทบ ‘วิกฤตเด็กไทยเกิดน้อย’ เขย่าสังคมไทย. แหล่งข้อมูล: https://tu.ac.th/thammasat-310165- crisis-thai-children-born-less. ค้นเมื่อวันที่ 15 มีนาคม 2566.

ปพิชญา กลางนอก. (2561). การประยุกต์ใช้เทคนิคแบบรวมเพื่อเพิ่มประสิทธิภาพของแบบจำลองตามกฎ. วิทยานิพนธ์วิทยาศาสตร์มหาบัณฑิต, มหาวิทยาลัยมหาสารคาม. มหาสารคาม. 53 หน้า.

รัชฎา เทพประสิทธิ์ และจรัญ แสนราช. (2563). การวิเคราะห์ปัจจัยที่มีผลต่อการเลือกสาขาวิชาของนักศึกษาระดับปริญญาตรี คณะครุศาสตร์ โดยใช้เทคนิคการทำเหมืองข้อมูล. วารสารบัณฑิตศึกษา มหาวิทยาลัยราชภัฏวไลยอลงกรณ์ ในพระบรมราชูปถัมภ์ 14(1): 134 - 146.

อัศวิน สุรวัชโยธิน และวรภัทร ไพรีเกรง. (2564). การสร้างตัวแบบการทำนายในการเลือกศึกษาต่อในระดับอุดมศึกษา โดยการใช้เทคนิคแบบบูรณาการในการแก้ปัญหาการจำแนกข้อมูลไม่สมดุลของกลุ่มผู้เรียน. วารสารวิทยาการและเทคโนโลยีสารสนเทศ 11(1): 65 – 79.

Bauer, E. and Kohavi, R. (1999). An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. Machine Learning 36: 105 – 139.

Breiman, L. (2001). Random Forest. Machine Learning 45: 5 - 32.

Chawla, V.N., Bowyer, W.K., Hall, O.L. and Kegelmeyer, P.W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 16(1): 321 - 357.

Hall, A.M. (1999). Correlation-based Feature Selection for Machine Learning. Thesis for the degree of Doctor of Philosophy, The University of Waikato. Hamilton, NewZealand. 198 pages.

Jin, X., Xu, A., Bie, R. and Guo, P. (2006). Machine Learning Techniques and Chi-Square Feature Selection for Cancer Classification Using SAGE Gene Expression Profiles. Data Mining for Biomedical Applications 3916: 106 - 115.

Lee, C. and Lee, G.G. (2006). Information gain and divergence-based feature selection for machine learning-based text categorization. Information Processing & Management 42(1): 155 - 165.

Natekin, A. and Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in Neurorobotics 7: 1 - 21.

Peng, J.Y.C., Lee, L.K. and Ingersoll, M.G. (2002). An Introduction to Logistic Regression Analysis and Reporting. The Journal of Educational Research 96(1): 3 - 14.

Quinlan, J.R. (1986). Induction of Decision Trees. Machine Learning 1: 81 - 106.

Rish, I. (2001). An Empirical Study of the Naïve Bayes Classifier. ResearchGate 3: 41 - 46.

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. Journal of Data Warehouse 5(4): 13 – 22.

Sripaoray, S. and Sinsomboonthong, S. (2017). Efficiency Comparison of Data Mining Classification Methods for Chronic Kidney Disease: A Case Study of a Hospital in India. Thai Science and Technology Journal 5(25): 839 - 853.

Vrigazova, B. (2021). The Proportion for Splitting Data into Training and Test Set for the Bootstrap in Classification Problems. Business Systems Research Journal 1(12): 228 - 242.

❑❑❑❑❑