การเพิ่มประสิทธิภาพการจำแนกข้อมูลการตัดสินใจเลือกซื้อเครื่องคอมพิวเตอร์โน้ตบุ๊คของนักศึกษามหาวิทยาลัยราชภัฏนครปฐมด้วยการปรับสมดุลข้อมูล

ศักดิ์สิทธิ์ แซ่ลิ้ม[1] , ไก้รุ่ง เฮงพระพรหม[1*] และ สุพจน์ เฮงพระพรหม[1]

[1]สาขาวิชาวิทยาการข้อมูล คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครปฐม
*wavetow1379@hotmail.com

บทคัดย่อ

การวิจัยนี้มีวัตถุประสงค์เพื่อปรับสมดุลข้อมูลและเปรียบเทียบเทียบประสิทธิภาพของแบบจำลองในการทำนายการตัดสินใจเลือกซื้อเครื่องคอมพิวเตอร์โน้ตบุ๊ค โดยใช้ 4 เทคนิคได้แก่ Decision Tree(J48) ,Naïve Bayes , k-nearest neighbors' algorithm (k-NN) และ multi-layer perceptron นำมาเปรียบเทียบประสิทธิภาพการจำแนกหาตัวแบบที่เหมาะสมในการทำนายการเลือกซื้อคอมพิวเตอ์โน้ตบุ๊คโดยใช้ข้อมูลที่เก็บรวบรวมมาของนักศึกษามหาวิทยาลัยราชภัฏนครปฐม จำนวน 101 คน โดยใช้ โปรแกรม Weka Version 3.8.6 เนื่องจาก ข้อมูลมีความไม่สมดุลจึงทำให้ได้ค่าความถูกต้องที่ต่ำมากไม่เหมาะสมที่จะนำไปใช้งาน ผู้วิจัยจึงแก้ปัญหาโดยการใช้วิธีสังเคราะห์ข้อมูลเพิ่มมาช่วยเพิ่มประสิทธิภาพ ผลลัพธ์ที่ได้เป็นที่พึงพอใจ พบว่าเทคนิค K-NN เป็นตัวแบบที่เหมาะสมที่สุดที่จะนำมาใช้จำแนกข้อมูลการตัดสินใจเลือกซื้อเครื่องคอมพิวเตอร์โน้ตบุ๊คของนักศึกษามหาวิทยาลัยราชภัฏนครปฐม ได้ค่าความถูกต้องสูงที่สุดอยู่ที่ 80% ค่าความแม่นยำ 0.79 ค่าระลึก 0.80

คำสำคัญ: การจำแนกข้อมูล ข้อมูลไม่สมดุล การปรับสมดุลข้อมูล

An Improving efficiency of data classification for notebook computer purchase decisions of Nakhon Pathom Rajabhat University students using data balancing

Saksit Salim[1], Kairung Hengpraprohm[1*], and Supojn Hengpraprohm[1]

[1]Program in data science, Faculty of Science and Technology, Nakhon Pathom Rajabhat University
*wavetow1379@hotmail.com

บทคัดย่อ

The purpose of this research is to balance the data and compare the efficiency of the model for predicting notebook computer purchase decisions by using 4 techniques: Decision Tree(J48), Naïve Bayes, k-nearest neighbors' algorithm (k- NN), and multi-layer perceptron. These technique are used to compare the performance of the appropriate classification model for predicting notebook computer purchases using the collected data of 101 Nakhon Pathom Rajabhat University students and the Weka software version 3.8.6. Due to the data is imbalanced, the accuracy is very low, it is not suitable for use. The researcher, therefore, solved the problem by using additional data synthesis methods to increase efficiency. The results are satisfactory. It is found that the K-NN technique is the most suitable model used to classify the purchasing decisions of notebook computers of Nakhon Pathom Rajabhat University students. Get the highest accuracy of 80%, a precision of 0.79, and a recall of 0.80.

Keywords: data classification, imbalanced data, data balancing

1. บทนำ

ปัจจุบันเทคโนโลยีมีบทบาทกับการใช้ชีวิตมากขึ้น คอมพิวเตอร์โน้ตบุ๊คมีความสำคัญในชีวิตประจำวัน เนื่องจากมีการเรียนการสอนในรูปแบบออนไลน์ และ มีการทำงานในรูปแบบออนไลน์มากขึ้น นักศึกษาต้องมีโน้ตบุ๊ค เพื่อศึกษาหรือพกพาเพื่อใช้ทำงานในสถานที่ต่างๆ และในปัจจุบันโน้ตบุ๊คก็มีหลายรุ่น หลายยี่ห้อ หลายราคา ทำให้การจะเลือกซื้อคอมพิวเตอร์โน้ตบุ๊คมีหลายปัจจัย ทั้งเรื่องของงบประมาณ และ สเปคต่างๆของเครื่องคอมพิวเตอร์โน้ตบุ๊ค ที่เหมาะสมกับแต่ละวัตถุประสงค์ของผู้ใช้งาน เพื่อให้ผู้ที่ซื้อไปแล้วพอใจในสินค้ามากที่สุด คุ้มค่ามากที่สุด และ เหมาะสมกับตัวเราที่สุดในการเลือกใช้งาน

จากการศึกษามุ่งเน้นที่จะศึกษาปัจจัยที่มีผลต่อการเลือกซื้อคอมพิวเตอร์โน้ตบุ๊คของนักศึกษา เนื่องจากนักศึกษาใช้คอมพิวเตอร์โน้ตบุ๊คเป็นจำนวนมาก โดยศึกษาจากกลุ่มตัวอย่างนักศึกษามหาวิทยาลัยราชภัฏนครปฐม ในการเลือกซื้อคอมพิวเตอร์โน้ตบุ๊ค ปัจจัยที่ทำให้ผู้ศึกษาเลือกที่จะ ศึกษากลุ่มตัวอย่างดังกล่าว เพื่อความสะดวกให้การเดินทางสำรวจและง่ายต่อผู้ตอบแบบสอบถาม และนำข้อมูลที่ได้ไปเปรียบเทียบประสิทธิภาพด้วย เทคนิค decision tree(J48), Naïve Bayes k-nearest neighbors’ algorithm (k-NN) และ multi-layer perceptron เพื่อหาค่าความถูกต้องในการจำแนกข้อมูล และนำไปตัดสินใจในการเลือกซื้อเครื่องคอมพิวเตอร์โน้ตบุ๊ค

ดังนั้นงานวิจัยนี้จึงต้องการศึกษาประสิทธิภาพของวิธีการแก้ปัญหาข้อมูลไม่สมดุลที่มีผลต่อการจำแนกผลข้อมูลการตัดสินใจเลือกซื้อเครื่องคอมพิวเตอร์โน้ตบุ๊คของนักศึกษามหาวิทยาลัยราชภัฏนครปฐม โดยใช้วิธีสังเคราะห์ข้อมูลเพิ่ม (Synthetic Minority Oversampling Technique : SMOTE) มาใช้เพื่อแก้ไขข้อมูลไม่สมดุลเพิ่มประสิทธิภาพในการจำแนกข้อมูลให้มีประสิทธิภาพสูงขึ้น และนำไปเปรียบเทียบประสิทธิภาพด้วยเทคนิคต่างๆเพื่อหาค่าความถูกต้อง

2. วัตถุประสงค์การวิจัย

2.1 เพื่อศึกษาเทคนิคในการจำแนกข้อมูลการเลือกซื้อเครื่องคอมพิวเตอร์โน้ตบุ๊คกรณีศึกษา นักศึกษามหาวิทยาลัยราชภัฏนครปฐม

2.2 เพื่อเปรียบเทียบประสิทธิภาพเทคนิคในการจำแนกข้อมูลการเลือกซื้อเครื่องคอมพิวเตอร์โน้ตบุ๊คกรณีศึกษานักศึกษามหาวิทยาลัยราชภัฏนครปฐม

2.3 เพื่อเพิ่มประสิทธิภาพการจำแนกข้อมูลการปรับสมดุลข้อมูลด้วย วิธีสังเคราะห์ข้อมูลเพิ่ม (Synthetic Minority Oversampling Technique: SMOTE)

3. วิธีดำเนินการวิจัย

3.1 กรอบแนวคิดในการวิจัย

กรอบแนวคิดในการวิจัย แสดงเป็นรูปภาพขั้นตอนการออกแบบการวิจัย แสดงดังภาพที่ 1

3.2 ข้อมูลสำหรับการวิจัย

ประชากร หมายถึง นักศึกษาที่ศึกษาอยู่ที่มหาลัยราชภัฏนครปฐม

กลุ่มตัวอย่าง หมายถึง นักศึกษามหาวิทยาลัยราชภัฏนครปฐม จำนวน 101 คน

ข้อมูลการศึกษาปัจจัยที่มีผลต่อการเลือกซื้อคอมพิวเตอร์โน้ตบุ๊คของนักศึกษามหาวิทยาลัยราชภัฏนครปฐม จำนวน 101 คน

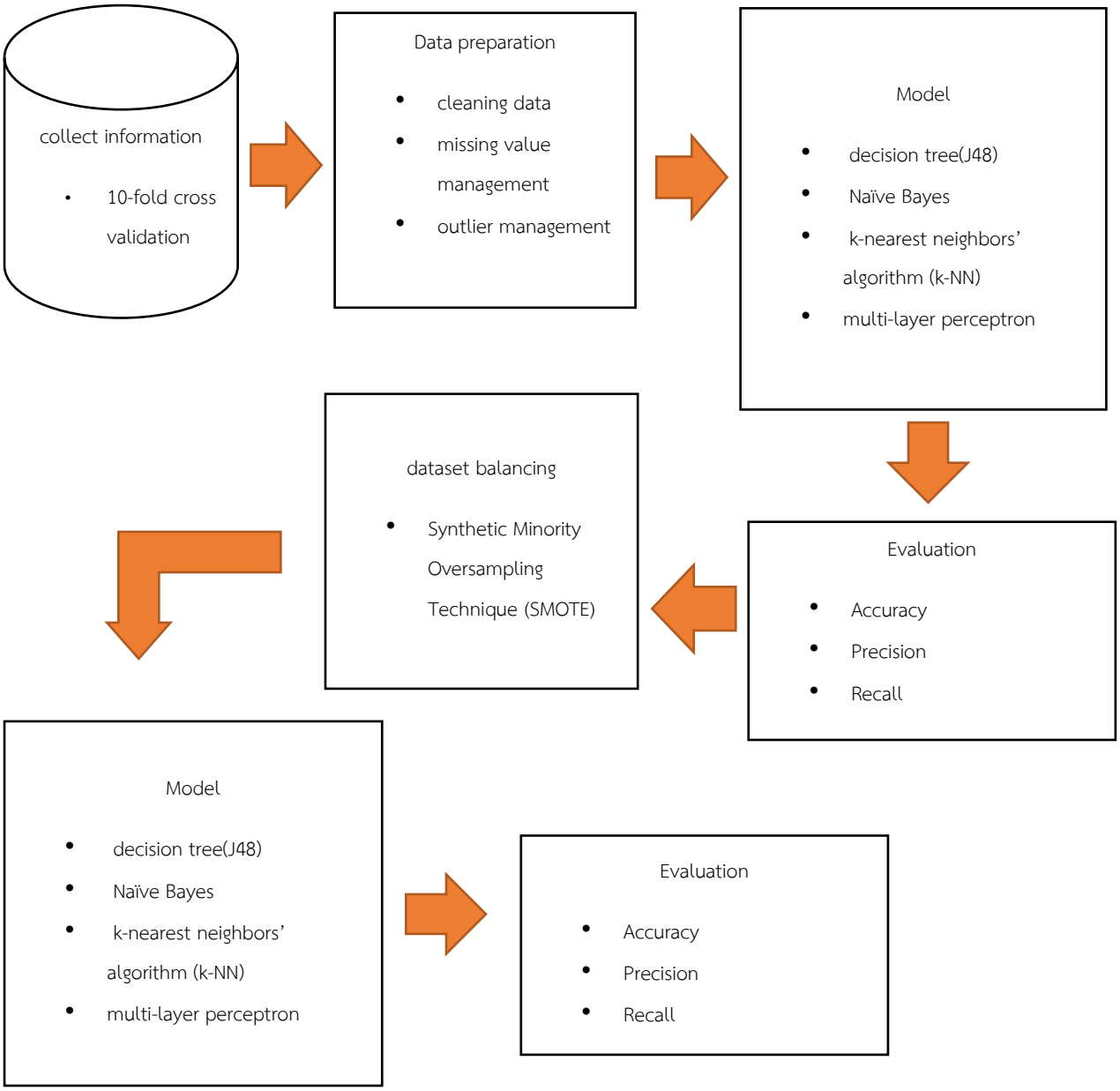
3.3 ขั้นตอนในการดำเนินการวิจัย (ขั้นตอนการทำ Data mining-KDD)

ขั้นตอนที่ 1 ทำความเข้าใจปัญหา (understand the problem)

เนื่องจากข้อมูลที่ได้จากแบบสอบถามปัจจัยที่มีผลต่อการเลือกซื้อคอมพิวเตอร์โน้ตบุ๊คของนักศึกษามหาวิทยาลัยราชภัฏนครปฐมมีความไม่สมดุลเนื่องจากAttributeบางตัวมีมากเกินไปและบางตัวมีน้อยเกินไปจึงทำให้ได้ค่าความถูกต้องที่ต่ำไม่เหมาะสมไปใช้งาน

ขั้นตอนที่ 2 ทำความเข้าใจข้อมูล (Data Understanding)

เก็บรวบรวมข้อมูลจากแบบสอบถามของนักศึกษามหาวิทยาลัยราชภัฏนครปฐมจาก google form ข้อมูลที่ใช้ในการสร้างแบบจำลองจะใช้ข้อมูลการศึกษาปัจจัยที่มีผลต่อการเลือกซื้อคอมพิวเตอร์โน้ตบุ๊คของนักศึกษามหาวิทยาลัยราชภัฏนครปฐม โดยมีข้อมูลทั้งหมด 101 ประกอบไปด้วย 30 แอตทริบิวต์ แสดงดังตารางที่ 1



ภาพที่ 1 ขั้นตอนการทำวิจัย เริ่มจากการเก็บข้อมูล เตรียมข้อมูล แบ่งเทรนแบ่งเทส สร้างโมเดล แล้วนำไปปรับสมดุลข้อมูล และสร้างโมเดล หาค่าความถูกต้อง

ตารางที่ 1 ตัวอย่างข้อมูลที่ใช้สำหรับวิเคราะห์

ข้อมูลทั่วไป	1.เพศ
	2.อายุ
	3.ชั้นปี
	4.วัตถุประสงค์
	5.ราคา
	6.การเลือกชำระเงิน
	7.จำนวนเดือนในการผ่อน
	8.เหตุจูงใจในการซื้อ
	9.รายได้ผู้ปกครอง
ปัจจัยในการเลือกซื้อด้านผลิตภัณฑ์	10.เป็นตราสินค้าที่ได้รับความนิยม
	11.ประสิทธิภาพมีความเหมาะสมกับการใช้งาน
	12.รูปแบบดีไซน์ รูปทรงของเครื่องที่ทันสมัย
	13.ความหลากหลายในการเลือกซื้อสินค้า
	14.ขนาดน้ำหนักสะดวกต่อการพกพา
ปัจจัยในการเลือกซื้อด้านราคาและการจัดจำหน่าย	15.ราคาเหมาะสมกับคุณภาพและอายุการใช้งานของโน้ตบุ๊ค
	16.มีหลายระดับราคาให้เลือกซื้อ
	17.มีการให้สินเชื่อ สามารถผ่อนชำระเป็นงวดๆ
	18.การลดราคาตามช่วงเทศกาลต่างๆ
	19.ราคาอุปกรณ์เสริมของคอมพิวเตอร์โน้ตบุ๊ค(เช่น เมาส์,คีย์บอร์ด,ลำโพง)
ปัจจัยในการเลือกซื้อด้านสถานที่	20.การเดินทางไปร้านค้ามีความสะดวกสบาย
	21.ความน่าเชื่อถือของตัวแทนจำหน่าย
	22.ตัวแทนจำหน่ายมีสินค้าให้ทดลองก่อนตัดสินใจ
	23.ร้านค้าที่ตั้งอยู่ในศูนย์รวม IT
	24.มีการขายสินค้าผ่านทางออนไลน์
ปัจจัยในการเลือกซื้อด้านการบริการ	25.การให้บริการของบริษัทจำหน่ายร้านค้า
	26.มีพนักงานที่สามารถให้คำแนะนำที่ดีแก่ลูกค้า
	27.มีบริการหลังการขายที่สะดวกและรวดเร็ว
	28.มีศูนย์ซ่อมคอมพิวเตอร์ในพื้นที่ใกล้เคียง
	29.มีบริการจัดส่งถึงหน้าบ้าน
	30.ยี่ห้อเครื่องคอมพิวเตอร์โน้ตบุ๊ค

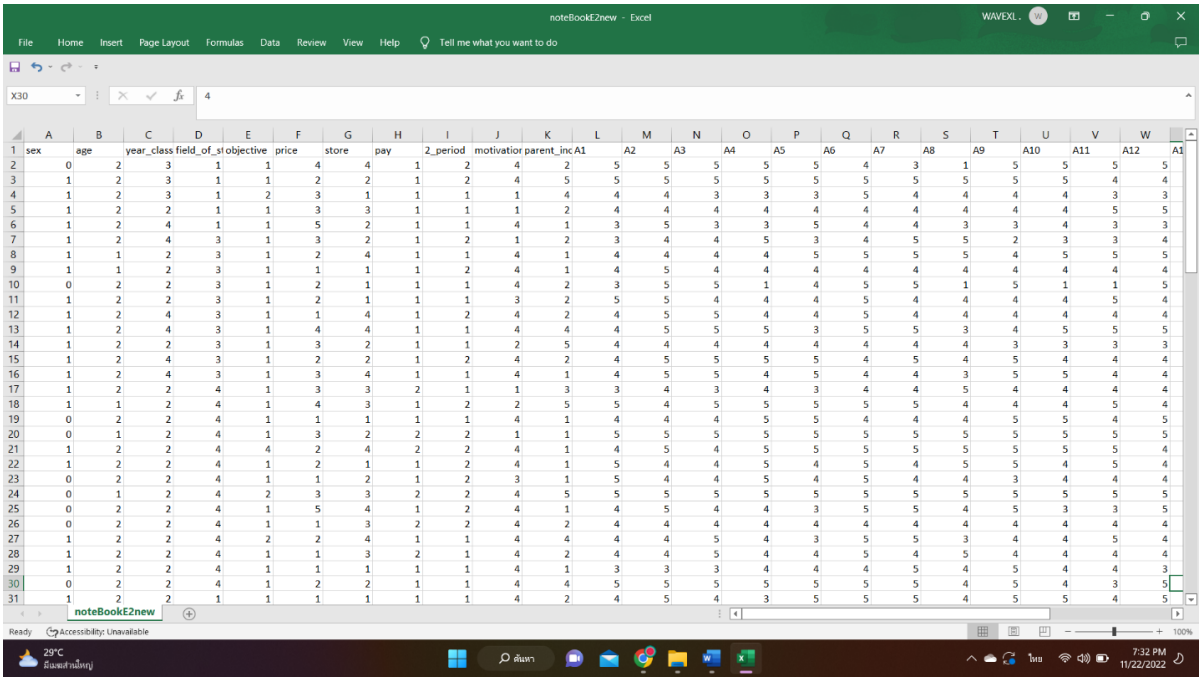
ขั้นตอนที่ 3 การเตรียมข้อมูล (Data Preparation)

เป็นการเตรียมข้อมูลโดยเตรียมข้อมูลดิบที่มีให้เป็นข้อมูลที่จะต้องใช้ในขั้นตอนที่เหลือตลอดจนเลือกตัวแปรที่ต้องการวิเคราะห์ให้เหมาะสมอีกทั้งแปลงรูปแบบของตัวแปรให้อยู่ใน รูปแบบเดียวกัน เพื่อให้ข้อมูลพร้อมสำหรับการนำไปสร้างแบบจำลอง ได้นำข้อมูลการศึกษาปัจจัยที่มีผลต่อการเลือกซื้อคอมพิวเตอร์โน้ตบุ๊คของนักศึกษามหาวิทยาลัยราชภัฏนครปฐม จำนวน 101 คน ข้อมูลแบ่งออกเป็น ขั้นตอนย่อย 3 ขั้นตอนดังนี้

1) ทำการคัดเลือกข้อมูล (Data Selection) เป็นเป็นการเลือกฟิลด์หรือข้อมูลที่ เกี่ยวข้อง ที่สนใจมาทำเหมืองข้อมูล และนำฟิลด์ที่ไม่สนใจออกโดยผู้วิจัยได้เลือกตัด ชั้นปี การเลือกชำระเงิน และ จำนวนเดือนในการผ่อน ออกเนื่องจากดูไม่เกี่ยวข้องกับปัจจัยในการเลือกซื้อเครื่องคอมพิวเตอร์โน้ตบุ๊ค

2) ทำการกรองข้อมูล (Data Cleaning) เป็นขั้นตอนการทำความสะอาดข้อมูล เพื่อจัดให้อยู่ในรูปแบบที่เหมาะสม จัดการข้อมูลขาดหาย (Missing data) ออกไป เช่น ข้อมูลที่มีค่า ว่าง หรือ ข้อมูลไม่ครบถ้วน ซึ่งเป็นข้อมูลที่ไม่สามารถนำมาวิเคราะห์ได้ หากนำมาวิเคราะห์ก็ จะทำให้ได้กฎความสัมพันธ์ที่ไม่ถูกต้อง

3) แปลงรูปแบบของข้อมูล (Data Transformation) เป็นการแปลงข้อมูลให้อยู่ใน รูปแบบที่พร้อมนำไปใช้ในการวิเคราะห์ขั้นตอนการแปลงข้อมูลในขั้นตอนการคัดเลือกให้เหมาะ สำหรับขั้นตอนการทำเหมืองข้อมูล โดยผู้วิจัยได้แปลงข้อมูลให้อยู่ในรูปของตัวเลขเหมือนกันเพื่อง่ายต่อการนำไปวิเคราะห์



ภาพที่ 2 การแปลงข้อมูลให้อยู่ในรูปแบบตัวเลขเหมือนกันเพื่อง่ายต่อการนำไปวิเคราะห์

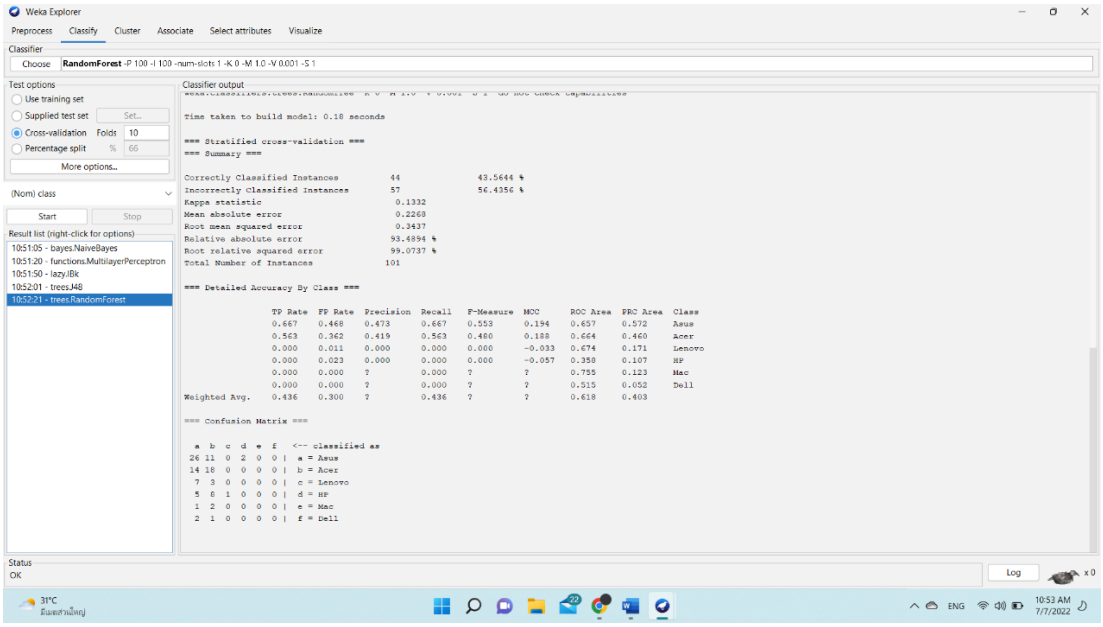
ขั้นตอนที่ 4 การสร้างแบบจำลอง (Modeling Phase)

โดยแบ่ง 10-fold cross validation และเลือกแบบจำลองที่เหมาะสม ในขั้นตอนนี้เลือกเทคนิคและขั้นตอนวิธีในการทำเหมืองข้อมูล ดังต่อไปนี้

1. Decision tree (J48)
2. Naïve Bayes
3. K-nearest neighbors’ algorithm (k-NN)
4. multi-layer perceptron

ขั้นตอนที่ 5 การประเมินผลแบบจำลอง (Evaluation Phase)

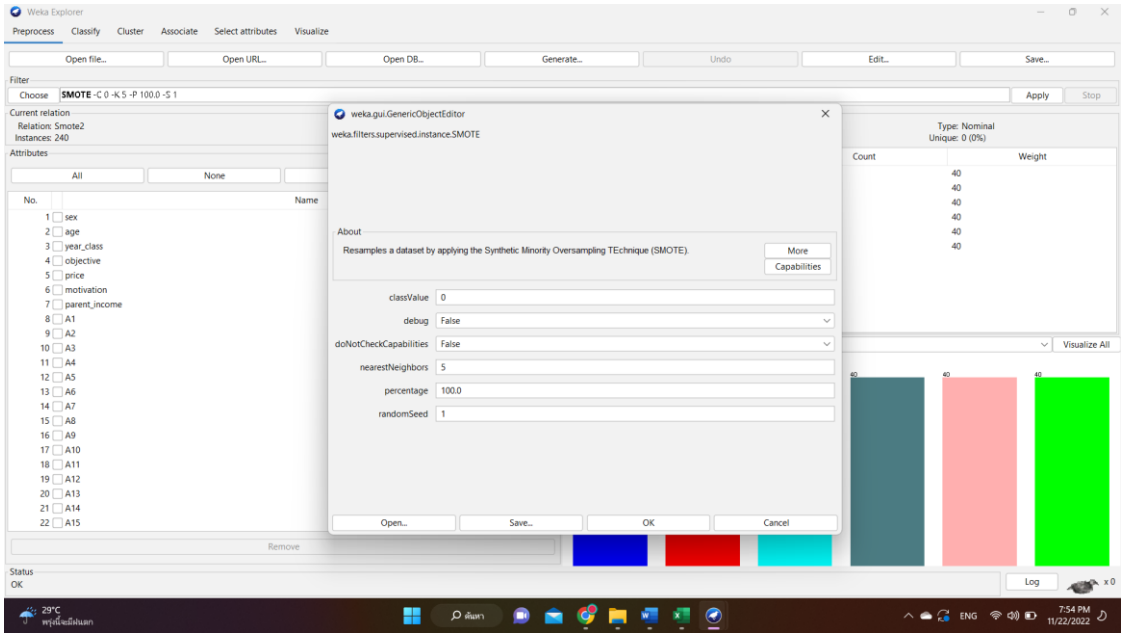
เป็นการประเมินแบบจำลองที่ใช้ในการวิเคราะห์ทั้งหมด เพื่อพิจารณาถึงความเหมาะสมใน การนำแบบจำลองไปประยุกต์ใช้ว่าแบบจำลองที่ได้มีความแม่นยำในการทำนายมากน้อย เพียงใด โดยมีการประเมินผลโมเดลที่เป็นค่าความถูกต้อง Accuracy , Precision และ Recall



ภาพที่ 3 การหาค่าความถูกต้อง Accuracy Precision และ Recall ก่อนการปรับสมดุลข้อมูล (dataset balancing)

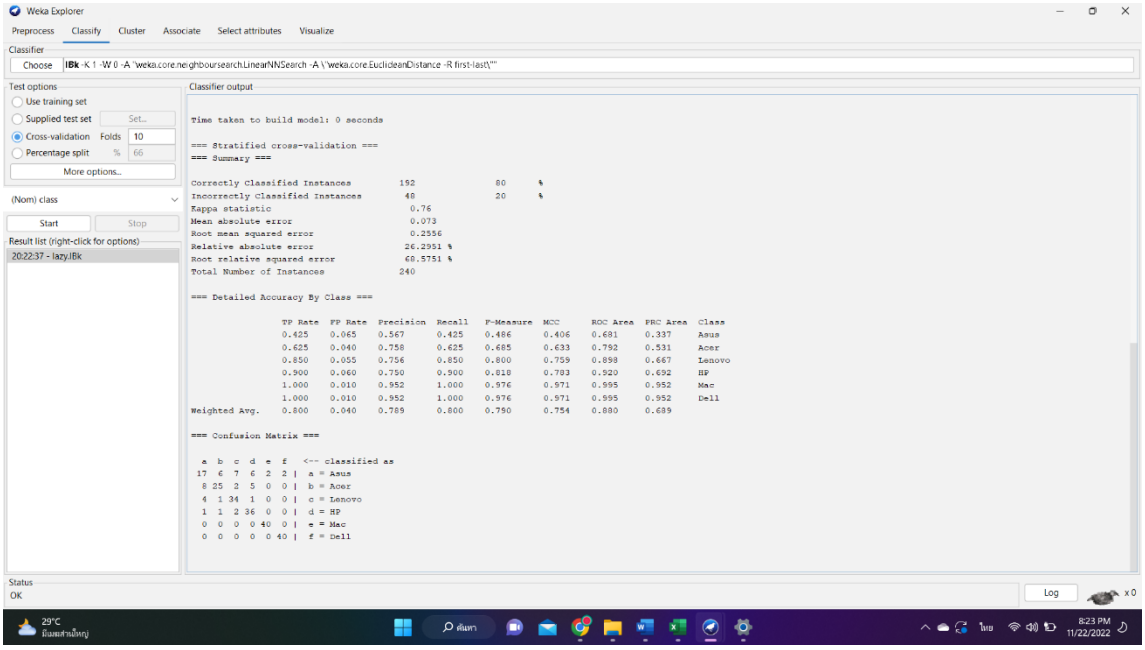
ขั้นตอนที่ 6 การปรับสมดุลข้อมูล (dataset balancing)

เนื่องจากข้อมูลไม่สมดุลจึงได้ค่าความถูกต้องต่ำ จึงเริ่มทำการเพิ่มประสิทธิภาพการจำแนกข้อมูลด้วยการปรับสมดุลข้อมูลโดยวิธีสังเคราะห์ข้อมูลเพิ่ม(Synthetic Minority Oversampling Technique : SMOTE)



ภาพที่ 4 การปรับสมดุลข้อมูล (dataset balancing)

โดยผู้วิจัยได้เลือกปรับสมดุลทีละ class Value และเลือกจำนวนเพื่อนบ้านของข้อมูลที่เหมาะสม และค่อยๆ เพิ่ม percentage ทำให้แต่ละคลาสเท่ากัน อยู่ที่คลาสละ 40 เพื่อให้ข้อมูลมีความสมดุลกันและนำไปวิเคราะห์ต่อไป



ภาพที่ 5 การหาค่าความถูกต้อง Accuracy Precision และ Recall หลังการปรับสมดุลข้อมูล (dataset balancing)

ขั้นตอนที่ 7 การนำไปใช้งาน (Deployment Phase)

หลังจากแก้ไขความไม่สมดุลของข้อมูลโดยวิธีสังเคราะห์ข้อมูลเพิ่ม(Synthetic Minority Oversampling Technique : SMOTE) ค่าความถูกต้อง Accuracy Precision และ Recall เพิ่มขึ้นอย่างมากจึงสามารถนำไปใช้พัฒนาต่อได้

3.4 การประเมิลผลการวิจัย

การวิจัยนี้ใช้วิธีการประเมินประสิทธิภาพโดยแบ่งข้อมูลสำหรับทดสอบด้วยวิธี 10-fold cross validation การประเมินประสิทธิภาพวัดได้จากค่าความถูกต้อง Accuracy Precision และ Recall

4. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

4.1 ทฤษฎีที่เกี่ยวข้อง

4.1.1 เหมืองข้อมูล

เหมืองข้อมูล (Data Mining) Jirayu Sitichai et al., [1] คือการค้นหาหรือการสกัดความรู้จากฐานข้อมูลขนาดใหญ่ กระบวนการที่กระทำกับข้อมูลจำนวนมาก เพื่อค้นหารูปแบบแนวทางและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น โดยอาศัยหลักการทางคณิตศาสตร์สถิติเพื่อนำความรู้ที่ได้นั้นมาใช้ในการแก้ปัญหา วางแผน หรือการดำเนินกลยุทธ์ขององค์กรให้ประสบความสำเร็จสูงสุด ซึ่งการวิเคราะห์ข้อมูลด้วยเทคนิคเหมืองข้อมูลสามารถแบ่งได้เป็น 2 ประเภทหลักๆ คือ

1.เทคนิคการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) จะเน้นที่การพิจารณาข้อมูลเป็นหลัก เช่นพิจารณาว่าข้อมูลมีความสัมพันธ์กันในลักษณะใดบ้าง เทคนิคในประเภทนี้จะแบ่งย่อยได้อีกคือ เทคนิคการค้นหากฎความสัมพันธ์ (Association Rule) และการแบ่งกลุ่มข้อมูล (Clustering)

2.เทคนิคการเรียนรูปแบบมีผู้สอน (Supervised Learning) เน้นการเรียนรู้จากข้อมูลที่มีอยู่ในอดีตเพื่อนำมาสร้างโมเดลสำหรับทำนายหรือคาดการณ์สิ่งที่จะเกิดขึ้นในอนาคต สามารถแบ่งย่อยได้อีกคือ การจำแนกประเภทข้อมูล

(Classification) และการประมาณค่าข้อมูล (Regression) ซึ่งทั้งสองเทคนิคจะมีความแตกต่างกันที่คำตอบที่ต้องการทำนาย ซึ่งการจำแนกประเภทข้อมูล (Classification) จะทำนายข้อมูลที่มีคำเป็น Nominalเช่น เพศชาย เพศหญิง หรือคำที่ไม่ใช่ตัวเลขนั่นเอง ส่วนการประมาณค่าข้อมูล (Regression) จะใช้กับข้อมูลคำตอบที่เป็นตัวเลขเท่านั้น

4.1.2 วิธีต้นไม้ตัดสินใจ (Decision Tree) เป็นการนำข้อมูลมาสร้างแบบจำลองการพยากรณ์ในรูปแบบโครงสร้างต้นไม้สามารถสร้างแบบจำลอง การจัดหมวดหมู่ได้จากกลุ่มตัวอย่างข้อมูลที่กำหนดไว้ล่วงหน้า และพยากรณ์กลุ่มของรายการที่ยังไม่เคยนำมาจัด หมวดหมู่ได้ด้วยรูปแบบของ Tree [1]

4.1.3 วิธีแบบเบย์ (Naïve Bayes) คือ การทำเหมืองข้อมูลที่ถูกสร้างขึ้นโดยหลักความน่าจะเป็น ซึ่งจะใช้การวิเคราะห์หาความน่าจะเป็น ของสิ่งที่ยังไม่เคยเกิดขึ้นด้วยการคาดเดาจากสิ่งที่เคยเกิดขึ้นมาก่อน โดยใช้ทฤษฎีของ Bayes ในการแก้ปัญหา [1]

4.1.4 วิธีเพื่อนบ้านใกล้สุด (k-NN) [2] เป็นวิธีที่ใช้ในการจัดแบ่งคลาส โดยเทคนิคนี้จะตัดสินใจว่า คลาสใดที่จะแทนเงื่อนไขหรือกรณีใหม่ๆ ได้บ้าง โดยการตรวจสอบจำนวนบางจำนวน (“K” ในขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด) ของกรณีหรือเงื่อนไขที่เหมือนกันหรือใกล้เคียงกันมากที่สุด โดยจะหาผลรวมของจำนวนเงื่อนไข หรือกรณีต่างๆ สำหรับแต่ละคลาส และกำหนดเงื่อนไขใหม่ๆ ให้คลาสที่เหมือนกันกับคลาสที่ใกล้เคียงกันมากที่สุด

4.1.6 วิธีเพอร์เซ็ปตรอนหลายชั้น (multi-layer perceptron) (Placeholder1)โครงข่ายประสาทเทียมแบบ MLP เป็นรูปแบบหนึ่งของโครงข่ายประสาทเทียมที่มีโครงสร้างเป็นแบบหลายๆชั้น ใช้สำหรับงานที่มีความซับซ้อนได้ผลเป็นอย่างดี โดยมีกระบวนการฝึกฝนเป็นแบบมีผู้สอน และใช้ขั้นตอนการส่งค่าย้อนกลับ สำหรับการฝึกฝนกระบวนการส่งค่าย้อนกลับประกอบด้วย 2 ส่วนย่อยคือ การส่งผ่านไปข้างหน้า (Forward Pass) การส่งผ่านย้อนกลับ (Backward Pass) สำหรับการส่งผ่านไปข้างหน้า ข้อมูลจะผ่านเข้าโครงข่ายประสาทเทียมที่ชั้นข้อ มูลเข้า และจะส่งผ่าน จากอีกชั้นหนึ่งไปสู่อีกชั้นหนึ่งจนกระทั่งถึงชั้นข้อมูลออก ส่วนการส่งผ่านย้อนกลับค่าน้ำหนักการเชื่อมต่อจะถูกปรับเปลี่ยนให้สอดคล้องกับกฎการแก้ข้อผิดพลาด (Error-Correction) คือผลต่างของผลตอบที่แท้จริง (Actual Response) กับผลตอบเป้าหมาย (Target Response) เกิดเป็นสัญญาณผิดพลาด (Error Signal) ซึ่งสัญญาณผิดพลาดนี้จะถูกส่งย้อนกลับเข้าสู่โครงข่ายประสาทเทียมในทิศทางตรงกันข้ามกับการเชื่อมต่อ และค่าน้ำหนักของการเชื่อมต่อจะถูกปรับจนกระทั่งผลตอบที่แท้จริงเข้าใกล้ผลตอบเป้าหมาย

4.1.7 วิธีสังเคราะห์ข้อมูลเพิ่ม (Synthetic Minority Oversampling Technique: SMOTE) เป็นเทคนิคการสุ่มตัวอย่างแบบพิเศษของการสุ่มเพิ่ม แทนที่จะสุ่มเพิ่มโดยใช้ข้อมูลเดิมแต่จะทำ การสังเคราะห์ข้อมูลขึ้นมาใหม่จากข้อมูลเดิมที่มีอยู่ หลักการเพื่อนบ้านที่อยู่ใกล้ที่สุดในการขยายขอบเขตการตัดสินใจของตัวแบบ

$$x_s = x_i + u \cdot (\hat{x}_i - x_i)$$

โดยที่ x_S แทน ข้อมูลที่สังเคราะห์ใหม่

x_i แทน ข้อมูลเดิมที่สุ่มมา

$\hat{x}_i$ แทน ข้อมูลที่เป็นเพื่อนบ้านของ x_i

u แทน ค่าสุ่มที่อยู่ระหว่าง 0 – 1

4.2 งานวิจัยที่เกี่ยวข้อง

Jirayu Sitichai et al., [1] โดย ใช้เทคนิคเหมืองข้อมูล 2 วิธีคือ วิธีต้นไม้ตัดสินใจ และวิธีแบบเบย์นำมาเปรียบเทียบประสิทธิภาพของโมเดลการ จำแนกหาตัวแบบที่ เหมาะสมเพื่อใช้ทำนายหาชนิดของไวน์โดยใช้ข้อมูล wine ของเว็บไซต์ UCI วิเคราะห์ข้อมูล บนพื้นฐานของวิธี 10-Fold Cross Validation โดยใช้โปรแกรม R studio ในการสร้างแบบจำลอง

Suphatsara Somjeta and Jaree Thongkam [3] มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของเทคนิคในเหมืองข้อมูลในการสร้างแบบจำลองจำแนกความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร งานวิจัยนี้ใช้6 เทคนิค ได้แก่ เทคนิคริปเปอร์เทคนิคต้นไม้ตัดสินใจแบบซี4.5 เทคนิคนาอีฟเบย์ เทคนิคซัพพอร์ตเวคเตอร์แมชชีนเทคนิคเคเนียร์เรสเนเบอร์และเทคนิคต้นไม้ป่าสุ่มมาสร้างแบบจำลองความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร โดยข้อมูลนั้นถูกรวบรวมมาเฉพาะความคิดเห็นของผู้ปกครองที่มีลักษณะเป็นข้อความภาษาไทยบนเครือข่ายสังคมออนไลน์ผ่านเว็บไซต์พันทิปและเฟซบุ๊คจำนวนทั้งหมด 1,925 ข้อความ

Nattavadee Hongboonmee and Praphasiri Trepanichkul [4] เปรียบเทียบประสิทธิภาพของการจำแนกข้อมูลด้วยอัลกอลิทึมเหมืองข้อมูลสามแบบ คือ โครงข่ายประสาทเทียม การเรียนรู้แบบเบย์และต้นไม้ตัดสินใจเพื่อให้ได้อัลกอลิทึมที่มีประสิทธิภาพสูงสุดที่จะถูกนำมาวิเคราะห์หาปัจจัยที่ส่งผลต่อความเสี่ยงการเกิดโรคไฮเปอร์ไทรอยด์โดยการลดการนำเข้าทีละปัจจัย ซึ่งข้อมูลที่นำมาใช้ในการทดลองเป็นข้อมูลจากโรงพยาบาลในจังหวัดพิษณุโลกจำนวน 323 ชุดข้อมูล ข้อมูลสำหรับการวิเคราะห์มีจำนวน 12 ปัจจัย

Phanthipa Phetchbunmee and Oranuch [2] เปรียบเทียบประสิทธิภาพอัลกอลิทึมเหมืองข้อมูลเพื่อจำแนกประเภทข้อมูลความสามารถทางการเรียนรู้ตามแนวทางพหุปัญญา สำหรับนักศึกษามหาวิทยาลัยเทคโนโลยีราชมงคลล้านนาตาก ทำการเก็บรวบรวมข้อมูลแบบสอบถามที่ผู้วิจัยได้พัฒนาขึ้น ตั้งแต่ปีการศึกษา 2558 – 2559 จำนวน 1,407 คน

Kamonrat Somjai [5] โดยมีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพโมเดลการตัดสินใจเลือกกลุ่มวิชาของนักศึกษาสาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยราชภัฏบุรีรัมย์ โดยใช้เหมืองข้อมูล 3 เทคนิควิธีคือ Decision Tree , Naïve Bayes และ Neural Network และใช้ข้อมูลนักศึกษาสาขาวิชาเทคโนโลยีสารสนเทศระหว่างปีการศึกษา 2555-2560 จำนวนทั้งสิ้น 407

Noppamas Akarachantachote and Direk Panitsupakamol [6] วัตถุประสงค์ของการวิจัยนี้ เพื่อเปรียบเทียบประสิทธิภาพการแก้ปัญหาข้อมูลไม่สมดุลด้วยการสุ่มตัวอย่างซ้ำระหว่างวิธีการสุ่มตัวอย่างเพิ่มข้อมูลเริ่มต้นอย่างสุ่ม และการสุ่มตัวอย่างเพิ่มกลุ่มส่วนน้อยด้วยการสังเคราะห์ เพื่อเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูลระหว่างการถดถอยลอจิสติก และต้นไม้ตัดสินใจ สำหรับการจำแนกกลุ่มรายได้ผู้ประกอบการร้านยาประเภท ข.ย.1 โดยค่าวัดประสิทธิภาพที่ใช้ในการเปรียบเทียบได้แก่ ค่าความแม่นยำ อัตราความถูกต้องในการทำนายกลุ่มส่วนน้อย อัตราความถูกต้องในการทำ นายกลุ่มส่วนมาก และค่าการวัดเอฟ

Kittisak Kerdprasop et al., [7] ในงานวิจัยได้ทำการทดลองกับข้อมูลจำนวน 3 ชุดข้อมูลใช้เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย เทคนิคการสุ่มลดตัวอย่างส่วนมาก และเทคนิคการสุ่มตัวอย่างซ้ำสำหรับการปรับปรุงข้อมูลที่ไม่สมดุล ใช้เทคนิคต้นไม้ตัดสินใจ คาร์ทเรนดอมฟรอเรส ซัพพอรท์เวกเตอร์แมชชีน และโครงข่ายประสาทเทียม ร่วมกับเทคนิครวมกลุ่มเอดาบูทและถุงจำแนก เพื่อสร้างแบบจำลองสำหรับจำแนกข้อมูลใช้วิธี 10-fold cross validation เพื่อวัดประสิทธิภาพของแบบจำลอง วัดค่าประสิทธิภาพด้วยค่าความแม่นยำ ค่าความระลึก และค่าเอฟ

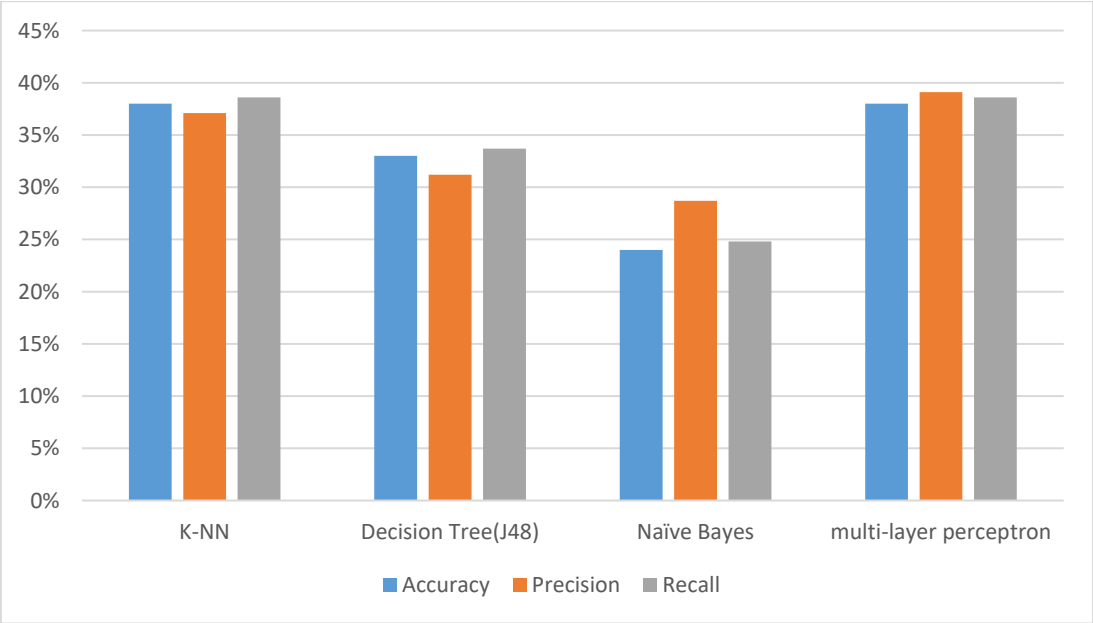
Wichit Lochirachunkol and Wichit Kesornsit [8] การจำแนกโดยใช้ข้อมูลที่ไม่สมดุลเป็นปัญหาสำคัญในเทคนิคการจำแนก ซึ่งการจำแนกข้อมูลที่มีข้อมูลในกลุ่มมากและกลุ่มน้อยปะปนกัน จะทำ ให้ข้อมูลในกลุ่มมากจะมีคุณสมบัติบางประการที่บดบังคุณสมบัติของกลุ่มน้อยทำให้การจำแนกข้อมูลในกลุ่มน้อยไม่สามารถจำแนกได้อย่างมีประสิทธิภาพ การวิจัยครั้งนี้จึงมีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีการแก้ปัญหาข้อมูลไม่สมดุลของข้อมูลผู้ป่วยโรคเบาหวานโดยการแก้ปัญหาข้อมูลไม่สมดุลที่ระดับข้อมูลจำนวน 4 วิธีคือ วิธีสุ่มเกิน วิธีสุ่มลด วิธีผสมผสาน และวิธีสังเคราะห์ข้อมูลใหม่ (SMOTE) โดยใช้เทคนิคการจำแนกคือวิธีการถดถอยโลจิสติกแบบมัลติโนเมียลและวิธีต้นไม้การตัดสินใจในการจำแนกผู้ป่วยโรคเบาหวาน

5. ผลการวิจัย

5.1 ตารางผลการเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูลการเลือกซื้อคอมพิวเตอร์โน้ตบุ๊คของนักศึกษามหาวิทยาลัยราชภัฏนครปฐม ก่อนทำ dataset balancing

ตารางที่ 2 ผลการเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูลการเลือกซื้อคอมพิวเตอร์โน้ตบุ๊คของนักศึกษามหาวิทยาลัยราชภัฏนครปฐม ก่อนทำ dataset balancing

เทคนิค	Accuracy	Precision	Recall
K-NN	38%	0.371	0.386
Decision Tree(J48)	33%	0.312	0.337
Naïve Bayes	24%	0.287	0.248
multi-layer perceptron	38%	0.391	0.386

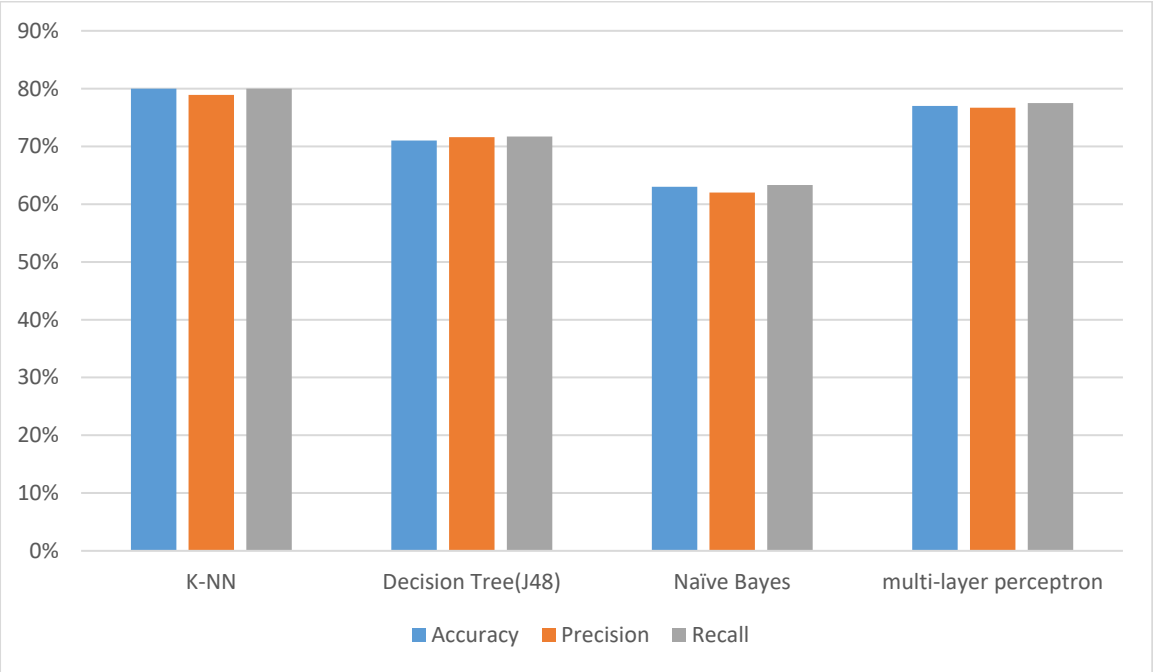


ภาพที่ 6 ผลการเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูลการเลือกซื้อคอมพิวเตอร์โน้ตบุ๊คของนักศึกษามหาวิทยาลัยราชภัฏนครปฐม ก่อนปรับสมดุลข้อมูล

ผลการทดลองจากตารางที่ 2 และภาพที่ 6 เทคนิคที่ให้ค่าความถูกต้องมากที่สุด คือ multi-layer perceptron Accuracy อยู่ที่ 38% Precision 0.391 Recall 0.386 และK-NN Accuracy อยู่ที่ 38% Precision 0.371 Recall 0.386 รองลงมา คือ Decision Tree(J48) Accuracy อยู่ที่ 3% Precision 0.312 Recall 0.337 ต่ำที่สุด คือ Naïve Bayes Accuracy อยู่ที่ 24% Precision 0.287 Recall 0.248

ตารางที่ 3 ผลการเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูลการเลือกซื้อคอมพิวเตอร์โน้ตบุ๊คของนักศึกษามหาวิทยาลัยราชภัฏนครปฐม หลังปรับสมดุลข้อมูล

เทคนิค	Accuracy	Precision	Recall
K-NN	80%	0.789	0.800
Decision Tree(J48)	71%	0.716	0.717
Naïve Bayes	63%	0.620	0.633
multi-layer perceptron	77%	0.767	0.775



ภาพที่ 7 ผลการเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูลการเลือกซื้อคอมพิวเตอร์โน้ตบุ๊คของนักศึกษามหาวิทยาลัยราชภัฏนครปฐม หลังปรับสมดุลข้อมูล

ผลการทดลองจากตารางที่ 3 และตารางที่ 4 และภาพที่ 7 เทคนิคที่ให้ค่าความถูกต้องมากที่สุด คือ K-NN Accuracy อยู่ที่ 80% Precision 0.789 Recall 0.800 และรองลงมา คือ multi-layer perceptron Accuracy อยู่ที่ 77% Precision 0.767 Recall 0.775 รองลงมา คือ Decision Tree (J48) Accuracy อยู่ที่ 71% Precision 0.716 Recall 0.717 ต่ำที่สุด คือ Naïve Bayes Accuracy อยู่ที่ 63% Precision 0.620 Recall 0.633

ตารางที่ 4 เปรียบเทียบประสิทธิภาพก่อนปรับสมดุลข้อมูลและหลังปรับสมดุลข้อมูลด้วยวิธีสังเคราะห์ข้อมูลเพิ่ม (Synthetic Minority Oversampling Technique : SMOTE)

เทคนิค	Accuracy	
	ก่อนปรับสมดุลข้อมูล	หลังปรับสมดุลข้อมูล
K-NN	38%	80%
Decision Tree (J48)	33%	71%
Naïve Bayes	24%	63%
multi-layer perceptron	38%	77%

6. สรุปผล

ผลการวิจัยครั้งนี้โดยพัฒนาและเปรียบเทียบตัวแบบการจำแนกทั้ง 4 เทคนิค ได้แก่ Decision Tree (J48), Naïve Bayes , k-nearest neighbors' algorithm (k-NN) และmulti-layer perceptron เนื่องจากข้อมูลไม่สมดุลจึงได้ค่าที่ต่ำมาก จึงแก้ปัญหาข้อมูลไม่สมดุลด้วยการใช้วิธีสังเคราะห์ข้อมูลเพิ่ม (Synthetic Minority Oversampling Technique : SMOTE) มาใช้ในการแก้ปัญหาข้อมูลไม่สมดุล ผลสรุปว่าวิธีสังเคราะห์ข้อมูลเพิ่ม (Synthetic Minority Oversampling Technique : SMOTE) แก้ปัญหาได้ดีมาก จึงทำให้ค่าความถูกต้อง (Accuracy) เพิ่มสูงขึ้นอย่างมาก จากผลการทดลอง พบว่าเทคนิคที่ให้ค่าความถูกต้องมากที่สุดคือ K-NN จากเดิม Accuracy อยู่ที่ 38% Precision 0.371 Recall 0.386 หลังปรับสมดุลข้อมูล Accuracy อยู่ที่ 80% Precision 0.789 Recall 0.800 และรองลงมา คือ multi-layer perceptron จากเดิม Accuracy อยู่ที่38% Precision 0.391 Recall 0.386 หลังปรับสมดุลข้อมูล Accuracy อยู่ที่ 77% Precision 0.767 Recall 0.775 รองลงมา คือ Decision Tree (J48) จากเดิม Accuracy อยู่ที่ 33% Precision 0.312 Recall 0.337 หลังปรับสมดุลข้อมูล Accuracy อยู่ที่ 71% Precision 0.716 Recall 0.717 ต่ำที่สุด คือ Naïve Bayes จากเดิม Accuracy อยู่ที่ 24% Precision 0.287 Recall 0.248 หลังปรับสมดุลข้อมูล Accuracy อยู่ที่ 63% Precision 0.620 Recall 0.633 สรุปได้ว่า K-NN เป็นตัวแบบที่เหมาะสมที่สุดที่จะนำมาใช้จำแนกข้อมูลการตัดสินใจเลือกซื้อเครื่องคอมพิวเตอร์โน้ตบุ๊คของนักศึกษามหาวิทยาลัยราชภัฏนครปฐม โดยใช้โปรแกรม Weka Version 3.8.6 เป็นโปรแกรมที่ใช้สร้างโมเดล

7.เอกสารที่เกี่ยวข้อง

[1] Jirayu Sitichai, Kairung Hengpraprohm and Supojn Hengpraprohm. (2021). A Comparison of Wine Quality Classification Performance by using Data Mining Techniques. The 13th NPRU National Academic Conference Nakhon Pathom Rajabhat University, P. 659-666, Nakhon Pathom. (In Thai)

[2] Oranuch, Lt. Colonel Phanthipa Phetchbunmee. (2017). A comparison efficiency of data mining algorithms for classification data of learning intellectual ability on multiple intelligences**. Proceedings of the 14th KU-KPS Conference**. P. 1091-1095. Nakhon Pathom. (In Thai)

[3] Suphatsara Somjeta and Jaree Thongkam. (2021). Performance Comparison of Data Mining Techniques for Building Classification Models of the Parent Toward Children who use Smart Phone. **Journal of Science and Technology, Ubon Ratchathani University**, Vol. 23, No. 1, P.21-30. (In Thai)

[4] Nattavadee Hongboonmee and Praphasiri Trepanichkul. (2019). Comparison of Data Classification Efficiency to Analyze Risk Factors that Affect the Occurrence of Hyperthyroid using Data Mining Techniques. **Journal of Information Science and Technology**. Vol.9, No.1, P.41-51. (In Thai)

[5] Kamonrat Somjai. (2020). Efficiency Comparison of Decision Making Models on Major Selection of Information Technology Students, Faculty of Science, Buriram Rajabhat University. The 5th Nation Science and Technology Conference. P. 2-8. Nakhon Si Thammarat Province. (In Thai)

[6] Noppamas Akarachantachote, Direk Panitsupakamol.(2019). **Comparison of Imbalanced Data Problem Solving for Income Classification of Type I Pharmacies Entrepreneur**. The 9th STOU National Research Conference. P.1582-1586. (In Thai)

[7] Kittisak Kerdprasop and Nittaya Kerdprasop, Kan Sritha. (2018). Comparison of sampling techniques for imbalanced data classification**. Journal of Applied Informatic and Technology**, Vol.1. No.1. P.20-37. (In Thai)

[8] Wichit Lochirachunkol, Wichit Kesornsit. (2018). Imbalanced Data Problem Solving in Classification of Diabetes Patients. **KKU Research Journal (Graduate Study)**.Vol.18, No.3, P.14-15. (In Thai)