บทความวิจัย

การสร้างตัวแบบทำนายการยืนยันสิทธิ์เข้าศึกษาของนักศึกษาใหม่ในระดับมหาวิทยาลัยช่วงสถานการณ์แพร่ระบาดของโควิด-19 กรณีศึกษา คณะวิทยาศาสตร์ มหาวิทยาลัยอุบลราชธานี

ไพชยนต์ คงไชย* สุภาวดี หิรัญพงศ์สิน และ ณัฏฐ์ ดิษเจริญ
ภาควิชาคณิตศาสตร์ สถิติ และคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยอุบลราชธานี

* ผู้นิพนธ์ประสานงาน โทรศัพท์ 08 6865 7476 อีเมล: phaichayon.k@ubu.ac.th DOI: 10.14416/j.kmutnb.2024.02.002
รับเมื่อ 6 พฤษภาคม 2565 แก้ไขเมื่อ 27 มิถุนายน 2565 ตอบรับเมื่อ 9 สิงหาคม 2565 เผยแพร่ออนไลน์ 19 กุมภาพันธ์ 2567

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อหาชุดข้อมูลที่เหมาะสมและเพื่อสร้างตัวแบบที่เหมาะสมสำหรับทำนายการรับเข้านักศึกษาใหม่ในระดับมหาวิทยาลัยช่วงสถานการณ์โควิด-19 ด้วยข้อมูลการรับเข้านักศึกษาใหม่ คณะวิทยาศาสตร์ มหาวิทยาลัยอุบลราชธานี ปีการศึกษา 2561–2564 จำนวน 4,748 ระเบียน และ 18 ปัจจัย เปรียบเทียบประสิทธิภาพการทำนายด้วย 5 อัลกอริทึม ได้แก่ แรนดอมฟอเรส นาอีฟเบย์ การถดถอยโลจิสติก ซัพพอร์ตเวกเตอร์ และอองซองค์เบิล และใช้ค่าความถูกต้องเป็นตัวชี้วัดในการประเมินประสิทธิภาพ การทดสอบแบ่งเป็น 2 ส่วน ได้แก่ การทดสอบที่ 1 เป็นการทดสอบหาชุดข้อมูลที่เหมาะสมสำหรับการสร้างตัวแบบเพื่อทำนายการยืนยันสิทธิ์เข้าศึกษาในสถานการณ์โควิด-19 พบว่า ชุดข้อมูลปีการศึกษา 2564 ซึ่งเป็นช่วงที่โควิด-19 ระบาดมาก มีค่าความถูกต้องมากที่สุดเท่ากับร้อยละ 66.67 และการทดสอบที่ 2 เป็นการทดสอบแบ่งชุดข้อมูลปีการศึกษา 2564 ตามจำนวนหลักสูตร ผลการทดลองพบว่า ประสิทธิภาพการทำนายมีค่าความถูกต้องสูงขึ้น โดยหลักสูตรชีววิทยามีค่าความถูกต้องสูงสุด คือ ร้อยละ 78.69 ด้วยอัลกอริทึมนาอีฟเบย์ นอกจากนี้ยังพบว่าอัลกอริทึมซัพพอร์ตเวกเตอร์สามารถทำนายได้ค่าความถูกต้องสูงสุดมากถึง 5 หลักสูตร เมื่อเปรียบเทียบกับอัลกอริทึมที่เหลือข้างต้น ดังนั้น การสร้างตัวแบบโดยการแบ่งข้อมูลตามจำนวนหลักสูตรจะมีประสิทธิภาพมากกว่าการสร้างหนึ่งตัวแบบจากการใช้ชุดข้อมูลทั้งหมด

คำสำคัญ: การทำเหมืองข้อมูลทางการศึกษา การเตรียมข้อมูล การจำแนกประเภทข้อมูล โควิด-19

การอ้างอิงบทความ: ไพชยนต์ คงไชย, สุภาวดี หิรัญพงศ์สิน และ ณัฏฐ์ ดิษเจริญ, "การสร้างตัวแบบทำนายการยืนยันสิทธิ์เข้าศึกษาของนักศึกษาใหม่ในระดับมหาวิทยาลัยช่วงสถานการณ์แพร่ระบาดของโควิด-19 กรณีศึกษา คณะวิทยาศาสตร์ มหาวิทยาลัยอุบลราชธานี," *วารสารวิชาการพระจอมเกล้าพระนครเหนือ*, ปีที่ 34, ฉบับที่ 2, หน้า 1–10, เลขที่บทความ 242-136040, เม.ย.–มิ.ย. 2567.

Research Article

Creating a Prediction Model for Prospective University Student Admissions During the COVID-19 Pandemic Situation: A Case Study of the Faculty of Science at Ubon Ratchathani University

Phaichayon Kongchai*, Supawadee Hiranpongsin and Nadh Ditcharoen
Department of Mathematics Statistics and Computers, Faculty of Science, Ubon Ratchathani University, Ubon Ratchathani, Thailand

* Corresponding Author, Tel. 08 6865 7476, E–mail: phaichayon.k@ubu.ac.th
DOI: 10.14416/j.kmutnb.2024.02.002
Received 6 May 2022; Revised 27 June 2022; Accepted 9 August 2022; Published online: 19 February 2024

Abstract

This research aimed to find a suitable dataset and to create a suitable model for predicting prospective university student admissions during the COVID-19 situation. The student admission information of Faculty of Science, Ubon Ratchathani University, during academic year 2018-2021 consisted of 4,748 records and 18 factors. To compare the predictive performance, five algorithms were used: Random Forest, Naïve Bayes, Logistic Regression, Support Vector Classification, and Ensemble. In addition, the accuracy metric was used for the performance evaluation. This experiment was divided into two parts. The first part of the experiment was to find a suitable data set to create a prediction model for prospective university student admissions during the COVID-19 situation. The results showed that the 2021 academic year dataset, which was during the COVID- 19 pandemic, had the highest accuracy at 66.67%. In the second part of the experiment, the 2021 data set was divided according to the number of courses. The results showed that the prediction performance had a higher accuracy value. The biology course had the highest accuracy of 78.69 percent using the Naïve Bayes algorithm. Furthermore, the support vector classification algorithm was able to predict the highest accuracy for up to 5 courses compared to other algorithms. Therefore, creating a model by dividing the data set according to the number of courses is more efficient than creating one model using the entire dataset.

Keywords: Educational Data Mining, Data Preparation, Classification, COVID-19

Please cite this article as: P. Kongchai, S. Hiranpongsin and N. Ditcharoen, "Creating a prediction model for prospective university student admissions during the COVID-19 pandemic situation: A case study of the faculty of science at Ubon Ratchathani university," *The Journal of KMUTNB*, vol. 34, no. 2, pp. 1–10, ID. 242-136040, Apr.–Jun. 2024 (in Thai).

1. บทนำ

งานรับเข้าศึกษาของมหาวิทยาลัยส่วนใหญ่ จะประกาศรับสมัครบุคคลเข้าศึกษาในมหาวิทยาลัยแต่ละรอบผ่านเว็บไซต์ของทางมหาวิทยาลัย และให้ผู้สมัครดำเนินการสมัครผ่านเว็บไซต์นั้น มหาวิทยาลัยอุบลราชธานีเป็นสถาบันการศึกษาหนึ่งที่มีการดำเนินการรับสมัครเช่นนี้ ผู้สมัครสามารถเลือกหลักสูตรของคณะต่าง ๆ และกรอกข้อมูลพื้นฐานเกี่ยวกับผลการศึกษาผ่านระบบ จากนั้นข้อมูลดังกล่าวจะถูกส่งต่อไปยังงานรับเข้าของแต่ละคณะต่อไป สำหรับคณะวิทยาศาสตร์มีหลักสูตรระดับปริญญาตรีทั้งหมด 10 หลักสูตร การรับเข้านักศึกษาใหม่ในแต่ละปีมีนักเรียนมาสมัครเข้าเรียนต่อจำนวนมาก และมีนักเรียนที่สละสิทธิ์เกือบร้อยละ 50 ของจำนวนผู้สมัครเช่นกัน ส่งผลให้งานรับเข้าของคณะต้องโทรติดตามสอบถามถึงเหตุผล ซึ่งได้คำตอบหลากหลายเหตุผล โดยเหตุผลส่วนใหญ่ คือ ลืมวันสอบสัมภาษณ์ หาห้องสอบสัมภาษณ์ไม่เจอ ลืมวันชำระเงิน ไม่มีเงินชำระค่าลงทะเบียน ทำให้งานรับเข้าต้องสูญเสียค่าใช้จ่ายและเวลาในการติดตามให้ครบทุกคน อีกทั้งปัญหาการกำหนดจำนวนรับเข้าในแต่ละรอบที่หลักสูตรต้องพิจารณาและกำหนดให้เหมาะสม เนื่องจากกระบวนการรับเข้ายังไม่สิ้นสุดวันที่ยืนยันสิทธิ์เป็นนักศึกษาใหม่ แต่ละหลักสูตรต้องกำหนดจำนวนรับเข้าในรอบถัดไปแล้ว ซึ่งหลักสูตรยังไม่ได้รับการยืนยันจำนวนนักศึกษาใหม่ที่ยืนยันสิทธิ์ ทำให้การกำหนดจำนวนรับเข้าอาจไม่สอดคล้องกับสถานการณ์ ณ ขณะนั้น ส่งผลให้นักเรียนที่กำลังยืนยันสิทธิ์ในรอบดังกล่าวเปลี่ยนใจไม่ยืนยันสิทธิ์ เนื่องจากนักเรียนทราบว่าจำนวนรับเข้าในรอบถัดไปรับจำนวนนักศึกษาใหม่ปริมาณที่มากขึ้น ในทางตรงกันข้าม ถ้ากำหนดจำนวนการรับเข้าในปริมาณที่น้อยลง จะส่งผลให้นักเรียนตัดสินใจไม่สมัครเนื่องจากโอกาสในการสอบติดจะลดลงทำให้ไปสมัครที่อื่นแทน นอกจากนี้ผู้สมัครสามารถตัดสินใจเลือกเข้าเรียนในสถาบันการศึกษาที่ใกล้หรือไกลจากภูมิลำเนาของผู้สมัครได้ง่ายขึ้น เนื่องจากสถานการณ์การแพร่ระบาดของโรคติดเชื้อไวรัสโคโรนาสายพันธุ์ใหม่ 2019 หรือ โควิด-19 สถาบันการศึกษาส่วนใหญ่มีการจัดการเรียนการสอนในรูปแบบออนไลน์ (Online) ใช้คอมพิวเตอร์ แท็บเล็ต หรืออุปกรณ์อื่น ๆ ที่สามารถเข้าถึงอินเทอร์เน็ต ผู้เรียนสามารถเรียนได้โดยไม่มีข้อจำกัดในเรื่องสถานที่ ดังนั้นการกำหนดจำนวนการรับเข้าที่สอดคล้องกับสถานการณ์ในปัจจุบัน รวมถึงการคาดการณ์การยืนยันสิทธิ์ของผู้สมัครจึงเป็นสิ่งจำเป็นต่อการบริหารจัดการ เช่น ด้านทรัพยากรการจัดการเรียนการสอน ด้านงบประมาณ และด้านบุคลากร

จากการศึกษางานวิจัยที่เกี่ยวข้องพบว่า มีงานวิจัย 2 ประเภท คือ งานวิจัยที่ใช้วิธีการเก็บข้อมูลจากแบบสอบถามเพื่อทำการวิเคราะห์ทางสถิติหาปัจจัยที่ส่งผลต่อการเลือกเข้าศึกษาต่อในระดับปริญญาตรี [1]–[3] และงานวิจัยที่ใช้วิธีการทำเหมืองข้อมูลเพื่อสร้างตัวพยากรณ์นักศึกษาใหม่ งานวิจัย [4] ได้เสนอวิธีการเปรียบเทียบประสิทธิภาพตัวแบบพยากรณ์จำนวนนักศึกษาใหม่ เพื่อวิเคราะห์ปัจจัยที่ส่งผลกับการเข้าศึกษาต่อในมหาวิทยาลัยราชภัฏนครปฐม โดยใช้เทคนิคอองซองค์เบิล (Ensemble) [5] ประมวลผลด้วยข้อมูลผู้สมัครที่ผ่านการคัดเลือกเป็นนักศึกษาใหม่ปีการศึกษา 2559–2560 แล้วทำการเปรียบเทียบค่าความถูกต้องของอัลกอริทึมโหวตอองซองค์เบิล (Vote Ensemble) อัลกอริทึมแบคกิ้ง (Bagging) และอัลกอริทึมแรนดอมฟอเรส (Random Forest) [6] ผลปรากฏว่าอัลกอริทึมแบคกิ้ง มีค่าความถูกต้องมากที่สุด ซึ่งสูงกว่า อัลกอริทึมโหวตอองซองค์เบิลและอัลกอริทึมแรนดอมฟอเรส งานวิจัย [7] ได้เสนอการศึกษาและพัฒนาตัวแบบพยากรณ์คุณลักษณะความเหมาะสมสำหรับการเลือกสมัครสาขาวิชาเรียนโดยใช้เทคนิคเหมืองข้อมูล โดยการจัดกลุ่มข้อมูลด้วยวิธีการเคมีนส์จากนั้นข้อมูลที่ได้ไปทำการหากฎความสัมพันธ์ด้วยอัลกอริทึมเอไพรออริ (Apriori) [8] แล้วประเมินความเหมาะสมของกฎความสัมพันธ์ด้วยผู้เชี่ยวชาญ ซึ่งผลการประเมินอยู่ในระดับมากที่สุด นอกจากนี้ได้สร้างตัวแบบพยากรณ์ด้วยอัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) ผลปรากฏว่า อัลกอริทึมแรนดอมฟอเรสมีค่าความถูกต้องสูงสุด และงานวิจัย [9] ได้เสนอการสร้างตัวแบบการทำนายในการเลือกศึกษาต่อในระดับอุดมศึกษา โดยการใช้เทคนิคแบบบูรณาการในการแก้ปัญหาการจำแนกข้อมูลไม่สมดุลของกลุ่มผู้เรียน ซึ่งได้แก้ปัญหาข้อมูลสูญหายด้วย

การใช้เทคนิคการแทนค่าสูญหายด้วยค่าเดียว (Single Imputation) และการปรับข้อมูลให้สมดุลด้วยเทคนิค SMOTE (Synthetic Minority Oversampling Technique) [10], [11] แล้วทำการทดลองสร้างตัวแบบด้วยอัลกอริทึมแบคกิ้งและสแตกกิ้ง (Stacking) ผลปรากฏว่าอัลกอริทึมสแตกกิ้ง ค่าความถูกต้องสูงสุด

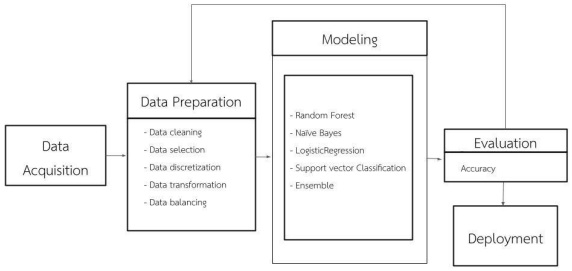
จากปัญหาดังกล่าวข้างต้น ผู้วิจัยจึงได้เสนอการสร้างตัวแบบทำนายการรับเข้านักศึกษาใหม่ในระดับมหาวิทยาลัยช่วงสถานการณ์แพร่ระบาดของโควิด-19 วัตถุประสงค์ของงานวิจัย คือ 1) เพื่อหาชุดข้อมูล (Data Set) ที่เหมาะสมสำหรับทำนายการรับเข้านักศึกษาใหม่ในสถานการณ์โควิด-19 และ 2) เพื่อสร้างตัวแบบที่เหมาะสมสำหรับทำนายการรับเข้านักศึกษาใหม่ในระดับมหาวิทยาลัยช่วงสถานการณ์แพร่ระบาดของโควิด-19 ซึ่งงานวิจัยนี้ผู้วิจัยมุ่งเน้นในการแก้ปัญหาด้วยวิธีการทำเหมืองข้อมูล เพราะสามารถนำผลที่ได้ไปประยุกต์ใช้เข้ากับระบบรับเข้าของมหาวิทยาลัยอุบลราชธานีได้ โดยตัวแบบที่ได้สามารถทำนายได้ว่าผู้สมัครคนใดจะยืนยันสิทธิ์หรือไม่ยืนยันสิทธิ์การเข้าเรียน ทั้งนี้เป็นการเพิ่มช่องทางสนับสนุนงานรับเข้าสำหรับคัดเลือกนักเรียนในการติดตาม และช่วยให้หลักสูตรสามารถกำหนดจำนวนการรับเข้าได้อย่างเหมาะสมและสอดคล้องกับสถานการณ์ที่อาจเปลี่ยนแปลงไป ทำให้สามารถบริหารจัดการงานด้านต่าง ๆ ได้อย่างมีประสิทธิภาพ

2. วัสดุ อุปกรณ์และวิธีการวิจัย

ผู้วิจัยได้ใช้จูปิเตอร์โน๊ตบุ๊ค (Jupyter Notebook) เป็นเครื่องมือในการเขียนโปรแกรมด้วยภาษาไพธอน เพื่อใช้ในการเก็บรวบรวมข้อมูล เตรียมข้อมูล และการวิเคราะห์ข้อมูล โดยมีตัวอย่างการเขียนโปรแกรมเพื่อสร้างตัวแบบดังรูปที่ 1 จากรูปเป็นฟังก์ชันในการสร้างตัวแบบโดยการใช้ 5 อัลกอริทึมในการประมวลผล ซึ่งผู้วิจัยได้อธิบายรายละเอียดของตัวแบบไว้ในขั้นตอนที่ 3 งานวิจัยนี้มีวิธีดำเนินการวิจัยประกอบด้วย 4 ขั้นตอน คือ 1) การเก็บรวบรวมข้อมูล (Data Acquisition) 2) การเตรียมข้อมูล (Data Preparation) 3) การสร้างตัวแบบ (Modeling) และ 4) การประเมินผลลัพธ์

```python
#สร้างตัวแบบ
def choose_model(x):
    if x==1:
        #อัลกอริทึม RandomForest
        print("RandomForestClassifier", end=" ")
        return RandomForestClassifier(n_estimators=1000)
    elif x==2:
        #อัลกอริทึม naive_bayes
        print("naive_bayes", end=" ")
        return GaussianNB()
    elif x==3:
        #อัลกอริทึม LogisticRegression
        print("LogisticRegression", end=" ")
        return LogisticRegression(random_state=1)
    elif x==4:
        #อัลกอริทึม Suppor Vector Machine
        print("SVC", end=" ")
        return make_pipeline(StandardScaler(), SVC(gamma='auto'))
    elif x==5:
        clf1 = RandomForestClassifier(n_estimators=1000)
        clf2 = GaussianNB()
        clf3 = make_pipeline(StandardScaler(), SVC(gamma='auto'))
        clf4 = LogisticRegression(random_state=1)
        print("Ensemble", end=" ")
        return VotingClassifier(estimators=[('rf', clf1),
        ('gnb', clf2), ('svc', clf3), ('lr', clf3)], voting='hard')
```

รูปที่ 1 ตัวอย่างการเขียนโปรแกรมเพื่อสร้างตัวแบบ



รูปที่ 2 กรอบแนวคิดขั้นตอนการพัฒนาตัวแบบ

(Evaluation) แสดงดังรูปที่ 2 โดยมีรายละเอียดในแต่ละขั้นตอนดังต่อไปนี้

ขั้นตอนที่ 1 การเก็บรวบรวมข้อมูล งานวิจัยนี้ได้เก็บรวบรวมข้อมูลมาจากงานรับเข้าคณะวิทยาศาสตร์ มหาวิทยาลัยอุบลราชธานี ตั้งแต่ปีการศึกษา 2561–2564 ซึ่งปีการศึกษา 2564 เป็นช่วงที่เกิดโควิด-19 ระบาดมาก ข้อมูลทั้งหมดจำนวน 5,760 ระเบียน และ 32 แอททริบิวต์ ประกอบด้วยข้อมูล ดังนี้ ลำดับผู้สมัคร รหัสผู้สมัคร รหัสประชาชน คำนำหน้าชื่อ ชื่อผู้สมัคร นามสกุลผู้สมัคร เกรดรวม สถานะการสมัคร ปีการศึกษาที่สมัคร ประเภทรอบการสมัคร (เช่น Quota, Portfolio, Admission) รอบ TCAS ที่สมัคร หมายเลขโทรศัพท์ อีเมล หมายเลข

โทรศัพท์มือถือ ชื่อบิดา นามสกุลบิดา หมายเลขโทรศัพท์ของบิดา ชื่อมารดา นามสกุลมารดา หมายเลขโทรศัพท์ของมารดา ชื่อโรงเรียน จังหวัด โควตาที่สมัคร เกรดภาษาไทย เกรดคณิตศาสตร์ เกรดวิทยาศาสตร์ เกรดสังคมศึกษา เกรดสุขศึกษา เกรดศิลปะ เกรดการงานอาชีพ เกรดภาษาต่างประเทศ และ สถานะผู้สมัคร “ยืนยัน” หรือ “สละสิทธิ์”

ขั้นตอนที่ 2 การเตรียมข้อมูล ผู้วิจัยได้ทำการเตรียมข้อมูลด้วยวิธีการดังต่อไปนี้

- การทำความสะอาดข้อมูล (Data Cleaning) ในขั้นตอนนี้ผู้วิจัยได้กำจัดข้อมูลสูญหาย (Missing Values) และข้อมูลที่มีค่าผิดพลาดทิ้งทั้งหมด

- การเลือกข้อมูล (Data selection) เพื่อเป็นการลดจำนวนแอททริบิวต์ ใช้เฉพาะแอททริบิวต์ที่มีความสำคัญจะเลือกแอททริบิวต์ที่เกี่ยวข้องทั้งหมด 18 แอททริบิวต์ และลบแอททริบิวต์ที่ไม่เกี่ยวข้องออก 14 แอททริบิวต์ ได้แก่ ลำดับผู้สมัคร รหัสผู้สมัคร รหัสประชาชน ชื่อผู้สมัคร นามสกุลผู้สมัคร หมายเลขโทรศัพท์ อีเมล์ หมายเลขโทรศัพท์มือถือ ชื่อบิดา นามสกุลบิดา หมายเลขโทรศัพท์ของบิดา ชื่อมารดา นามสกุลมารดา และ หมายเลขโทรศัพท์ของมารดา

- การแบ่งช่วงข้อมูล (Data Discretization) เนื่องจากข้อมูลเกรดเป็นค่าต่อเนื่องตั้งแต่ 00.00–04.00 เพื่อเพิ่มโอกาสในการเกิดรูปแบบของข้อมูล ผู้วิจัยได้ทำการแบ่งช่วงข้อมูลให้เป็นค่าไม่ต่อเนื่อง โดยปรับข้อมูลเกรดทั้งหมดให้เป็น 7 ช่วงข้อมูล ดังนี้ A (เกรด >= 3.50) B (เกรด >= 3.00) C (เกรด >= 2.50) D (เกรด >= 2.00) E (เกรด >= 1.50) F (เกรด >= 1.00) และ Z (เกรด >= 0.00)

- การแปลงข้อมูล (Data Transformation) ในขั้นตอนนี้ผู้วิจัยได้ทำการปรับขนาดของข้อมูลด้วยการทำนอร์มัลไลเซชัน (Normalization) เพื่อลดขนาดข้อมูลและลดทรัพยากรในการประมวลผลข้อมูล อีกทั้งผู้วิจัยได้ทำการแปลงข้อมูลให้อยู่ในรูปแบบของตัวเลขด้วยการทำ One Hot Encoding [12] เพื่อให้สามารถประมวลผลกับอัลกอริทึมที่เลือกมาสร้างตัวแบบได้

- การปรับข้อมูลให้สมดุล (Data Balancing) หลังจากการทำความสะอาดข้อมูล พบว่า ข้อมูลมีความต่างของ Label ในอัตรา 2 ต่อ 1 โดยมีทั้งหมดจำนวน 4,748 ระเบียน ได้แก่ “เลือกเข้าเรียน (ยืนยันสิทธิ์)” 2,958 ระเบียน และ “ไม่เลือกเข้าเรียน (ไม่ยืนยันสิทธิ์)” 1,790 ระเบียน และจากงานวิจัย [9]–[11] แสดงให้เห็นว่าการปรับสมดุลข้อมูลด้วยขั้นตอนวิธี SMOTE ทำให้ผลการทดสอบได้ค่าความถูกต้องที่สูง ดังนั้นผู้วิจัยจึงได้ทำการปรับข้อมูลให้สมดุลด้วยเทคนิคการทำ SMOTE (Over Sampling) ดังนั้นจำนวนข้อมูลหลังปรับสมดุลมีจำนวน 5,916 ระเบียน 18 แอททริบิวต์ ตัวอย่างคุณลักษณะข้อมูลงานรับเข้าแสดงดังตารางที่ 1

ขั้นตอนที่ 3 การสร้างตัวแบบ จากการศึกษางานวิจัยที่เกี่ยวข้องของผู้วิจัย ส่งผลให้งานวิจัยนี้ได้เลือกใช้อัลกอริทึมที่เป็นที่นิยมในการเปรียบเทียบประสิทธิภาพในการสร้างตัวแบบทั้งหมด 5 อัลกอริทึม ได้แก่ Random Forest (RF), Naïve Bayes (NB), Logistic Regression (LR), Support Vector Classification (SVC), และ Ensemble (ENS) ด้วยค่าพารามิเตอร์ที่กำหนดไว้แล้ว (Default Parameter) ในแต่ละอัลกอริทึม โดยมีหลักการทำงานดังต่อไปนี้

อัลกอริทึม Random Forest ใช้หลักการของต้นไม้ตัดสินใจ (Decision Tree) ในการสร้างตัวแบบ โดยจะสร้างต้นไม้ตัดสินใจจำนวน N ต้น ด้วยมาตรวัด Gini Index ดังสมการที่ (1) จากนั้นนำผลลัพธ์จากการทำนายของแต่ละต้นมาคิดเป็นคำตอบของตัวแบบ โดยจะเลือกผลโหวตที่มากที่สุด [6]

$$Gini(t) = 1 - \sum_{j=1}^{n} p_j^{\,2} \tag{1}$$

โดยที่ t แทนข้อมูลสำหรับฝึก และเป็นความถี่ของคลาส j

อัลกอริทึม Naïve Bayes ใช้หลักการความน่าจะเป็นด้วยทฤษฎีของเบย์ โดยทำการหาความสัมพันธ์ของระหว่างตัวแปร เพื่อใช้ในการสร้างตัวแบบ [12] ดังสมการที่ (2)

$$P(C_i \mid X) = \frac{P(X \mid C_i)P(C_i)}{P(X)} \tag{2}$$

ตารางที่ 1 คุณลักษณะของข้อมูลงานรับเข้า ปีการศึกษา 2561 – 2564

ลำดับ	แอททริบิวต์	คำอธิบาย	ตัวอย่างข้อมูล
1	PREFIXNAME	คำนำหน้าชื่อ	นาย, นางสาว
2	ENTRYGPAX	เกรดรวม	A, B, C, D, E, F, Z
3	APPLICANTSTATUS	สถานะการสมัคร	10, 16, 17, 19, 30, 31, 32, 33, 35, 40, 41, 43, 45, 46, 47, 48, 49, 50
4	ACADYEAR	ปีการศึกษาที่สมัคร	2561, 2562, 2563, 2564
5	APPLICANTTYPENAME	ประเภทรอบการสมัคร	Quota1, Quota2, Quota2, Portfolio1, Portfolio2, Portfolio3, Portfolio4, Portfolio5, Admission1, Admission2, Direct Admission
6	รอบ TCAS	รอบ TCAS ที่สมัคร	1, 2, 3, 4, 5
7	SCHOOLNAME	ชื่อโรงเรียน	นารีนุกูล, เขมราฐพิทยาคม, ... (721), อำนาจเจริญ
8	PROVINCENAME	จังหวัด	มหาสารคาม, อุบลราชธานี, ... (67), กาญจนบุรี
9	QUOTANAME	โควตาที่สมัคร	คณิตศาสตร์, จุลชีววิทยา, ... (7), วิทยาการข้อมูลและนวัตกรรมซอฟต์แวร์
10	ภาษาไทย	เกรดภาษาไทย	A, B, C, D, E, F, Z
11	คณิตศาสตร์	เกรดคณิตศาสตร์	A, B, C, D, E, F, Z
12	วิทยาศาสตร์	เกรดวิทยาศาสตร์	A, B, C, D, E, F, Z
13	สังคมศึกษา	เกรดสังคมศึกษา	A, B, C, D, E, F, Z
14	สุขศึกษา	เกรดสุขศึกษา	A, B, C, D, E, F, Z
15	ศิลปะ	เกรดศิลปะ	A, B, C, D, E, F, Z
16	การงานอาชีพ	เกรดการงานอาชีพ	A, B, C, D, E, F, Z
17	ภาษาต่างประเทศ	เกรดภาษาต่างประเทศ	A, B, C, D, E, F, Z
18	สถานะผู้สมัคร (Target)	สถานะผู้สมัคร	ยืนยันสิทธิ์, ไม่ยืนยันสิทธิ์

โดยที่ $P(C_i \mid X)$ คือ ค่าความน่าจะเป็นที่เกิดแอททริบิวต์ X ก่อนแล้วเป็นคลาส C_i, $P(C_i)$ คือ ค่าความน่าจะเป็นในการเกิดคลาส และ $P(X)$ คือ ค่าความน่าจะเป็นในการเกิดแอททริบิวต์ X

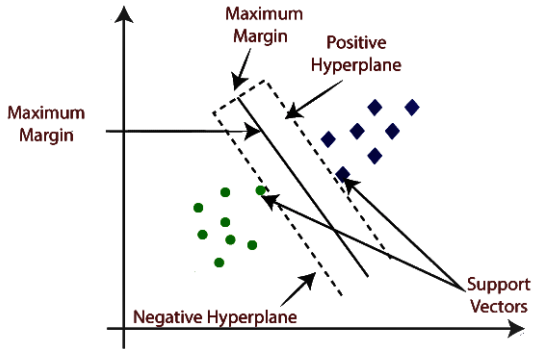
อัลกอริทึม Logistic Regression ใช้หลักการทางสถิติที่วิเคราะห์สมการแบบถดถอย เพื่อทำนายโอกาสที่จะเกิดเหตุการณ์ที่สนใจ แล้วใช้ Hypothesis Function เพื่อทำให้สามารถจำแนกข้อมูลได้ [13]

$$\log\left(\frac{P(y)}{Q(y)}\right) = \log\left(\frac{P(y)}{1-P(y)}\right) = b_0 + b_1 x_1 + \ldots b_n x_n \quad (3)$$

โดยที่ $P(y)$ แทนความน่าจะเป็นในการเกิดเหตุการณ์ที่สนใจ $Q(y)$ แทนความน่าจะเป็นของการไม่เกิดเหตุการณ์ที่สนใจ b แทนค่าสัมประสิทธิ์การถดถอย และ x แทนแอททริบิวต์ ซึ่งการประมาณค่าสัมประสิทธิ์การถดถอยใช้ความน่าจะเป็นสูงสุด หรือ Maximum Likelihood เมื่อเทียบผลการทำนาย สมการจะถดถอยเพื่อหาค่าทำนายของตัวแปรให้ใกล้เคียงกับข้อมูลจริงมากที่สุด

อัลกอริทึม Support Vector Classification ใช้หลักการหาสัมประสิทธิ์ของสมการ เพื่อสร้างเส้นแบ่งแยกข้อมูล (Hyperplane) โดยมีผลรวมระยะห่าง (Margin) ของเส้นตรงที่เป็นเส้นแบ่ง เส้นที่แบ่งกลุ่มที่กว้างมากที่สุดของทั้งสองกลุ่มเรียกว่า Maximum Margin แต่ข้อมูลส่วนใหญ่เป็นข้อมูลแบบไม่เชิงเส้น จึงมีการนำเคอร์เนลฟังก์ชันมาใช้งาน เพื่อให้สามารถจำแนกข้อมูลบนระนาบได้หลายมิติ และเวกเตอร์ที่อยู่ข้างระนาบจะเรียกว่าเวกเตอร์สนับสนุน (Support Vectors) [14] ดังรูปที่ 3

อัลกอริทึม Ensemble ใช้หลักการรวมหลายอัลกอริทึมเพื่อเพิ่มประสิทธิภาพในการจำแนกข้อมูล โดยการสร้างแบบจำลองเพียงแบบเดียว โดยแต่ละอัลกอริทึมที่นำมา



รูปที่ 3 แนวความคิดของอัลกอริทึม SVC

รวมกันนั้นจะถูกโหวตเพื่อตัดสินสุดท้าย เพื่อให้ได้ผลลัพธ์หนึ่งเดียวที่ใช้เป็นคำตอบของการจำแนกข้อมูล [15] โดยงานวิจัยนี้ได้มัดรวมอัลกอริทึม คือ Random Forest, Naïve Bayes, Logistic Regression (LR) และ Support Vector Classification

ขั้นตอนที่ 4 การประเมินผลลัพธ์ งานวิจัยนี้ได้ทำการเปรียบเทียบประสิทธิภาพของตัวแบบที่สร้างขึ้นด้วยการใช้เกณฑ์ค่าความถูกต้อง (Accuracy) ซึ่งแสดงผลเป็นร้อยละ และได้แบ่งข้อมูลสำหรับการทดสอบเป็นข้อมูลสำหรับฝึกปริมาณร้อยละ 70 ของข้อมูลทั้งหมด และข้อมูลสำหรับทดสอบปริมาณร้อยละ 30 ของข้อมูลทั้งหมด

3. ผลการทดลอง

งานวิจัยนี้ได้แบ่งการทดลองออกเป็น 2 การทดสอบ โดยมีรายละเอียดของแต่ละการทดสอบดังต่อไปนี้

3.1 การทดสอบที่ 1

เป็นการทดสอบหาชุดข้อมูลที่เหมาะสมสำหรับการสร้างตัวแบบเพื่อทำนายการรับเข้านักศึกษาใหม่ในสถานการณ์โควิด-19 การทดสอบนี้ได้แบ่งชุดข้อมูลสำหรับฝึก (Training Data Set) เป็น 5 ชุดข้อมูลตามปีการศึกษาที่รับเข้า และชุดข้อมูลสำหรับทดสอบ (Testing Data Set) เพียงปีเดียว คือ ชุดข้อมูลปีการศึกษา 2564 (ปีที่เกิดโควิด-19) ผลการทดสอบแสดงดังตารางที่ 2 ซึ่งผลการทดสอบแสดงให้เห็นว่าชุดข้อมูลฝึกที่มีค่าความถูกต้องในการสร้างตัวแบบสูงที่สุด คือ ชุดข้อมูลปีการศึกษา 2564 โดยมีค่าความถูกต้องถึงร้อยละ 66.67

ตารางที่ 2 ผลการทดสอบค่าความถูกต้องของการหาชุดข้อมูลที่เหมาะสมสำหรับการสร้างตัวแบบ

อัลกอริทึม	ชุดข้อมูลสำหรับฝึก (ปีการศึกษาที่รับเข้า)				
	ปี 61	ปี 62	ปี 63	ปี 64	ปี 61–63
RF	64.21*	64.21	60.38*	66.67*	62.96
NB	50.73	64.43*	45.72	56.02	63.25
LR	61.04	61.92	57.73	58.46	61.78
SVC	63.25	62.00	60.16	59.86	62.89
ENS	62.29	62.00	58.46	59.69	63.33*
ค่าสูงสุด	64.21	64.43	60.38	**66.67***	63.33

หมายเหตุ: * แทนค่าความถูกต้องสูงสุดของข้อมูลสำหรับฝึก

3.2 การทดสอบที่ 2

เป็นการทดสอบการสร้างตัวแบบด้วยการแบ่งแยกข้อมูลตามหลักสูตรด้วยชุดข้อมูลปีการศึกษา 2564 เนื่องจากผลการทดสอบที่ 1 ดังกล่าวข้างต้นที่ชุดข้อมูลฝึกปีการศึกษา 2564 มีค่าความถูกต้องสูงที่สุด การทดสอบที่ 2 นี้ จะแบ่งข้อมูลเป็น 10 ชุดข้อมูลตามจำนวนหลักสูตร คือ 1) ฟิสิกส์ชีวการแพทย์ 2) เทคโนโลยีสารสนเทศและการสื่อสาร 3) จุลชีววิทยา 4) วิทยาการข้อมูลและนวัตกรรมซอฟต์แวร์ 5) เคมี 6) อาชีวอนามัยและความปลอดภัย 7) ชีววิทยา 8) วิทยาศาสตร์สิ่งแวดล้อม 9) คณิตศาสตร์ และ 10) เทคโนโลยียางและพอลิเมอร์

ผลการทดสอบที่ 2 แสดงดังตารางที่ 3 ซึ่งผลการทดสอบแสดงให้เห็นว่า อัลกอริทึม SVC สามารถทำนายได้ค่าความถูกต้องสูงสุดมากถึง 5 หลักสูตร ได้แก่ ฟิสิกส์ชีวการแพทย์ วิทยาการข้อมูลและนวัตกรรมซอฟต์แวร์ วิทยาศาสตร์สิ่งแวดล้อม คณิตศาสตร์ และ เทคโนโลยียางและพอลิเมอร์ โดยหลักสูตรที่มีค่าความถูกต้องในการทำนายมากกว่าหรือเท่ากับร้อยละ 66.67 (ค่าความถูกต้องจากการทดลองที่ 1) มีจำนวน 6 หลักสูตร ได้แก่ ฟิสิกส์ชีวการแพทย์ เทคโนโลยีสารสนเทศและการสื่อสาร วิทยาการข้อมูลและนวัตกรรมซอฟต์แวร์ ชีววิทยา วิทยาศาสตร์สิ่งแวดล้อม และคณิตศาสตร์

ตารางที่ 3 ผลการทดสอบค่าความถูกต้องของการทดสอบการสร้างตัวแบบด้วยการแบ่งแยกข้อมูลตามหลักสูตร

ลำดับ	หลักสูตร	อัลกอริทึม (ร้อยละความถูกต้อง)					ค่าสูงสุด
		RF	NB	LR	SVC	ENS	
1	ฟิสิกส์ชีวการแพทย์	61.76	29.41	**75.00***	**75.00***	63.24	75.00
2	เทคโนโลยีสารสนเทศและการสื่อสาร	57.89	42.11	60.53	60.53	**68.42***	68.42
3	จุลชีววิทยา	63.93*	54.10	59.02	59.02	62.30	63.93
4	วิทยาการข้อมูลและนวัตกรรมซอฟต์แวร์	57.14	26.19	64.29	**66.67***	59.52	66.67
5	เคมี	56.90	58.62	50.00	56.90	**60.34***	60.34
6	อาชีวอนามัยและความปลอดภัย	46.34	**51.22***	48.78	48.78	48.78	51.22
7	ชีววิทยา	72.13	**78.69***	68.85	68.85	72.13	78.69
8	วิทยาศาสตร์สิ่งแวดล้อม	65.22	21.74	**69.57***	**69.57***	56.52	69.57
9	คณิตศาสตร์	72.09	39.53	**76.74***	**76.74***	69.77	76.74
10	เทคโนโลยียางและพอลิเมอร์	44.44	**55.56***	**55.56***	**55.56***	**55.56***	55.56
จำนวนความถี่ของอัลกอริทึมที่มีค่าความถูกต้องมากที่สุด		**1**	**3**	**4**	**5**	**3**	**78.69**

หมายเหตุ: * แทนค่าความถูกต้องสูงสุดของอัลกอริทึมของแต่ละหลักสูตร

โดยหลักสูตรชีววิทยามีค่าความถูกต้องมากที่สุดคือ ร้อยละ 78.69 และมีเพียง 4 หลักสูตรที่มีค่าความถูกต้องในการทำนายน้อยกว่า 66.67 ได้แก่ จุลชีววิทยา เคมี อาชีวอนามัยและความปลอดภัย และเทคโนโลยียางและพอลิเมอร์

ดังนั้น การสร้างตัวแบบที่เหมาะสมสำหรับทำนายการรับเข้านักศึกษาใหม่ในสถานการณ์โควิดของคณะวิทยาศาสตร์ มหาวิทยาลัยอุบลราชธานี ข้อมูลที่นำมาสร้างตัวแบบควรเป็นข้อมูลที่ถูกแบ่งกลุ่มแยกตามหลักสูตร ตามผลของการทดสอบที่ 2) โดยใช้ชุดข้อมูลปีการศึกษา 2564 ซึ่งเป็นช่วงที่โควิด-19 ระบาดมากที่สุด และเลือกใช้อัลกอริทึม Support Vector Classification (SVC) ด้วยการทดสอบที่แบ่งชุดข้อมูลเป็นข้อมูลสำหรับฝึกและข้อมูลสำหรับทดสอบปริมาณร้อยละ 70 และร้อยละ 30 ของข้อมูลทั้งหมดตามลำดับ ซึ่งจะทำให้การทำนายมีประสิทธิภาพที่มีความแม่นยำสูงขึ้น

4. อภิปรายผลและสรุป

งานวิจัยนี้มีวัตถุประสงค์เพื่อหาชุดข้อมูลและเพื่อสร้างตัวแบบทำนายการรับเข้านักศึกษาใหม่ในสถานการณ์โควิดของคณะวิทยาศาสตร์ มหาวิทยาลัยอุบลราชธานี ข้อมูลที่นำมาใช้ในงานวิจัยเป็นข้อมูลจากงานรับเข้าตั้งแต่ปีการศึกษา 2561–2564 ซึ่งปีการศึกษา 2564 เป็นช่วงที่โควิด-19 ระบาดมาก การทดสอบแบ่งเป็น 2 การทดสอบ ได้แก่ การทดสอบที่ 1 เป็นการทดสอบเพื่อหาชุดข้อมูลที่เหมาะสมสำหรับการสร้างตัวแบบเพื่อทำนายการรับเข้าในสถานการณ์โควิดของคณะวิทยาศาสตร์ มหาวิทยาลัยอุบลราชธานี และการทดสอบที่ 2 เป็นการทดสอบการสร้างตัวแบบด้วยการแบ่งแยกข้อมูลตามโดยใช้ชุดข้อมูลที่ได้จากผลการทดสอบที่ 1 การทดสอบจะแบ่งข้อมูลออกเป็นข้อมูลฝึกร้อยละ 70 และข้อมูลทดสอบร้อยละ 30 ของข้อมูลทั้งหมด อัลกอริทึมที่นำมาใช้ในการทดสอบประกอบด้วย 5 อัลกอริทึม ได้แก่ Random Forest (RF), Naïve Bayes (NB), Logistic Regression (LR), Support Vector Classification (SVC), และ Ensemble (ENS) ซึ่งบางอัลกอริทึมถูกนำมาใช้สอดคล้องกับงานวิจัย [7] และ [9]–[11] ผลการทดสอบพบว่า ชุดข้อมูลที่เหมาะสมสำหรับนำมาสร้างตัวแบบ คือ ชุดข้อมูลของปีการศึกษา 2564 โดยมีค่าความถูกต้องมากที่สุด คือ ร้อยละ 66.67 เนื่องจากข้อมูลที่นำมาทดสอบมีการแบ่งเป็นข้อมูลฝึกและข้อมูลทดสอบแยกจากกัน ดังนั้นผลการทดสอบการหาชุดข้อมูลที่ได้จึงไม่มีความลำเอียง และข้อมูลที่นำมาสร้างตัวแบบควรเป็นข้อมูลที่ถูกแบ่งกลุ่มแยกตามหลักสูตร ซึ่งสอดคล้องกับงานวิจัย [9] ที่มีการทดสอบการ

จำแนกข้อมูลโดยแยกตามกลุ่มสายวิชา ทำให้การทำนายมีประสิทธิภาพที่มีความแม่นยำสูงขึ้น โดยหลักสูตรชีววิทยามีค่าความถูกต้องสูงสุดคือร้อยละ 78.69 ด้วยการใช้อัลกอริทึม NB แม้ว่าการเลือกใช้อัลกอริทึม SVC จะได้ค่าความถูกต้องไม่สูงมากเท่าอัลกอริทึม NB แต่สามารถทำนายได้ค่าความถูกต้องสูงสุดมากถึง 5 หลักสูตร ดังนั้นจากผลการวิจัยนี้คณะผู้วิจัยจะนำไปประยุกต์ใช้เข้ากับระบบงานรับเข้าของมหาวิทยาลัยอุบลราชธานี เพื่อเป็นการเพิ่มช่องทางสนับสนุนงานรับเข้าสำหรับคัดเลือกนักเรียนในการติดตาม และช่วยให้หลักสูตรสามารถกำหนดจำนวนการรับเข้าได้อย่างเหมาะสมและสอดคล้องกับสถานการณ์ที่อาจเปลี่ยนแปลงไปได้อย่างมีประสิทธิภาพ งานวิจัยในอนาคตคณะผู้วิจัยเพิ่มการเก็บรวบรวมคุณลักษณะข้อมูลที่เกี่ยวข้องของจำนวนยอดผู้ติดเชื้อโควิด จำนวนผู้ได้ฉีดวัคซีน ข้อมูลผู้ปกครอง เช่น รายได้และอาชีพ และประยุกต์ใช้เทคนิคอื่น ๆ ของการเรียนรู้ของเครื่องเพื่อสร้างตัวแบบทำนายนี้ให้มีความแม่นยำที่สูงขึ้น

5. กิตติกรรมประกาศ

งานวิจัยนี้ได้รับเงินอุดหนุนการวิจัยจากมหาวิทยาลัยอุบลราชธานี ความเห็นในรายงานผลการวิจัยนี้เป็นของผู้วิจัย ผู้ให้ทุนไม่จำเป็นต้องเห็นด้วยเสมอไป และขอขอบคุณคณะวิทยาศาสตร์ มหาวิทยาลัยอุบลราชธานี ที่สนับสนุนการดำเนินงานวิจัยนี้

เอกสารอ้างอิง

[1] B. Srisombat, “Factors affecting the grouping decision making to continue higher education study in the faculty of science and technology pibulsongkram rajabphat university,” *Journal of Humanities and Social Sciences of Pibulsongkram Rajabhat University*, vol. 16, no. 2, pp. 189–200, 2015 (in Thai).

[2] S. Srisontisuk, M. buasri, and P. Wanong, “Factors affecting the decision on bachelor degree study at khon kaen university of the students through the thai university central admission system (TCAS System), in the academic year 2019,” *Journal of School of Administrative Studies Academic*, vol. 3, no. 3, pp. 33–47, 2020 (in Thai).

[3] P. Nittayakamolphun and T. Kaewkwankrai, “Factors affecting learning behavior of students studying statistics for economist,” *Journal of Humanities and Social Sciences University of Phayao*, vol. 9, no. 2, pp. 213–237, 2021 (in Thai).

[4] P. Jittawitsutigul and J. Sanrat, “The comparison of model efficiency to forecast the number of new students for analysis of factors affecting on admission to nakhon pathom rajabhat university by using ENSEMBLE,” in *Proceedings Nakhon Pathom Rajabhat University*, 2017 (in Thai).

[5] J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso, “Rotation forest: A new classifier ensemble method,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 10, pp. 1619–1630, 2006.

[6] M. Pal, “Random forest classifier for remote sensing classification,” *International Journal of Remote Sensing*, vol. 26, no. 1, pp. 217–222, 2005.

[7] S. Wanon and R. Muangsan, “A study and development of forecasting model for the suitability characteristics on the applying major selection by using data mining techniques,” *Journal of Management Sciences Suratthani Rajabhat University*, vol. 7, no. 1, pp. 135–152, 2020 (in Thai).

[8] R. Agrawal, T. Imieliński, and A. Swami, “Mining

association rules between sets of items in large databases," in *Proceedings the 1993 ACM SIGMOD International Conference on Management of Data*, 1993, pp. 207–216.

[9] A. Surawatchayotin and W. Paireekreng, "The predictive model of higher education guidance using integrated techniques for imbalanced data of learner groups," *Journal of Information Science and Technology*, vol. 11, no. 1, pp. 65–74. 2021 (in Thai).

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and P. W. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[11] W. Prachuabsupakij, T. Boonyoung, and S. Boonsri, "A construction model for predicting student's academic achievement by smartphone usage behaviors using data mining techniques," *The Journal of KMUTNB*, vol. 31, no. 3, pp. 550–560, 2021 (in Thai).

[12] I. Rish, "An empirical study of the naive Bayes classifier, in *Proceedings the International Joint Conference on Artificial Intelligence*, vol. 3, no. 22, pp. 41–46, 2001.

[13] E. Antipov and E. Pokryshevskaya, "Applying CHAID for logistic regression diagnostics and classification accuracy improvement," *Journal of Targeting, Measurement and Analysis for Marketing*, vol. 18, no. 2, pp. 109–117, 2010.

[14] L. Zhang, W. Zhou, and L. Jiao, "Wavelet support vector machine," *Journal of IEEE Transactions on Systems, Man, and Cybernetics*, vol. 34, no. 1, pp. 34–39, 2004.

[15] W. N. Street and Y. Kim, "A streaming ensemble algorithm (SEA) for large-scale classification," in *Proceedings of the Seventh Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 2001, pp. 377–382.