การใช้เทคโนโลยีการเรียนรู้ของเครื่อง เพื่อทำนายผลการเรียนของนักเรียน

STUDENTS' PERFORMANCE PREDICTION USING MACHINE LEARNING

ขวัญตา ศิลปไพบูลย์พานิช

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

2562

การใช้เทคโนโลยีการเรียนรู้ของเครื่อง เพื่อทำนายผลการเรียนของนักเรียน

ขวัญตา ศิลปไพบูลย์พานิช

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2562

STUDENTS' PERFORMANCE PREDICTION USING MACHINE LEARNING

KHWANTA SINLAPAPHAIBOONPHANIT

A Master's Project Submitted in Partial Fulfillment of the Requirements

for the Degree of MASTER OF SCIENCE

(Information Technology)

Faculty of Science, Srinakharinwirot University

2019

สารนิพนธ์

เรื่อง

การใช้เทคโนโลยีการเรียนรู้ของเครื่อง เพื่อทำนายผลการเรียนของนักเรียน

ของ

ขวัญตา ศิลปไพบูลย์พานิช

ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

ของมหาวิทยาลัยศรีนครินทรวิโรฒ

.....

(รองศาสตราจารย์ นายแพทย์ฉัตรชัย เอกปัญญาสกุล)

คณบดีบัณฑิตวิทยาลัย

.....

คณะกรรมการสอบปากเปล่าสารนิพนธ์

..... ที่ปรึกษาหลัก

(ผู้ช่วยศาสตราจารย์ ดร.จันตรี ผลประเสริฐ)

..... ประธาน

(ดร.สุทธิพงศ์ ธัชยพงษ์)

..... กรรมการ

(อาจารย์ ดร.ศิริสรรพ เหล่าหะเกียรติ)

ชื่อเรื่อง	การใช้เทคโนโลยีการเรียนรู้ของเครื่อง เพื่อทำนายผลการเรียนของนักเรียน
ผู้วิจัย	ขวัญตา ศิลปไพบูลย์พานิช
ปริญญา	วิทยาศาสตรมหาบัณฑิต
ปีการศึกษา	2562
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร. จันตรี ผลประเสริฐ

การทำนายผลการเรียนของนักเรียนในสาขาวิชาวิศวกรรมศาสตร์หรือทางวิทยาศาสตร์และเทคโนโลยีเป็นหนึ่งในหัวข้อวิจัยที่น่าสนใจในช่วง 4-5 ปีที่ผ่านมา งานวิจัยในอดีตใช้แบบจำลองการเรียนรู้ของเครื่อง เช่น K-Nearest Neighbor หรือ Decision Trees โดยวิเคราะห์จากผลการเรียนวิชาอื่นๆหรือจากแบบสอบถาม อย่างไรก็ตามวิธีดังกล่าวยังมีประสิทธิภาพไม่มากเท่าที่ควร เนื่องมาจากสาเหตุหลายประการ อาทิ การใช้ข้อมูลจากการตอบแบบสอบถามซึ่งขาดความน่าเชื่อถือหรือชุดข้อมูลมีจำนวนจำกัด งานวิจัยนี้นำเสนอการใช้เทคโนโลยีการเรียนรู้ของเครื่องในการทำนายผลการเรียนรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษ โดยงานวิจัยนี้มุ่งหวังที่จะศึกษาและเปรียบเทียบสมรรถนะของแบบจำลองการเรียนรู้ของเครื่องจำนวน 5 แบบจำลอง ได้แก่ Random Forest, Logistic Regression, Support Vector Machine, K-Nearest Neighbor และ Extreme Gradient Boosting โดยใช้ข้อมูลผลการเรียนในอดีตของนักเรียนระดับชั้นมัธยมศึกษาตอนต้น จากผลการทดลองพบว่าแบบจำลองที่มีประสิทธิภาพเฉลี่ยสะสมดีที่สุดในการทำนายผลการเรียนรายวิชาคณิตศาสตร์ คือแบบจำลอง K-Nearest Neighbor ที่ค่า accuracy 69.25% macro f1 67.00% รายวิชาวิทยาศาสตร์คือแบบจำลอง Extreme Gradient Boosting ที่ค่า accuracy 72.75% macro f1 67.00% และรายวิชาภาษาอังกฤษคือแบบจำลอง Logistic Regression ที่ค่า accuracy 64.00% macro f1 61.00% โดยจากผลการทดลองพบว่าโมเดลยังไม่สามารถทำนายผลการเรียนของผู้เรียนที่อยู่ในระดับพอใช้หรือระดับดีได้แม่นยำนักและควรจะมีการใช้ข้อมูลอื่นๆ อาทิ เช่น ข้อมูลแบบสอบถามของผู้เรียน ข้อมูลการมาเรียน หรือข้อมูลของผู้สอนมาเสริมเพื่อเพิ่มประสิทธิภาพของโมเดลให้แม่นยำยิ่งขึ้น

คำสำคัญ : เทคนิคการเรียนรู้ของเครื่อง, การทำนายผลการเรียน, การศึกษา

Title	STUDENTS' PERFORMANCE PREDICTION USING MACHINE LEARNING
Author	KHWANTA SINLAPAPHAIBOONPHANIT
Degree	MASTER OF SCIENCE
Academic Year	2019
Thesis Advisor	Assistant Professor Dr. Chantri Polprasert

The prediction of student grades in engineering, science or technology has become one of the main research problems due to the need to identify the factors contributing to the academic performance of students. Existing grade prediction systems which employ traditional machine learning models, such as K-nearest neighbor (KNN), decision trees exploiting past academic performance and answers from the questionnaires to predict grades yielded a poor performance. In this work, machine learning models were used to exploit past academic grades to predict performance for subjects in mathematics, science and English. The performance of Random Forest, Logistic Regression, Support Vector Machine, KNN and Extreme Gradient Boosting (XGBoost) models investigated, compared and predicted the performances of the students. From the experiment, the proposed model employed KNN achievements of 69.25% accuracy and 67.00% macro f1 when predicting performances in mathematics, XGBoost achieved 72.75% accuracy and 67.00% macro f1 when predicting performances in science and Logistic Regression achieved 64.00% accuracy and 61.00% macro f1 when predicting the performances in English. The model exhibited marginal performance when predicting students with poor or average academic performance, but this could be due to limited datasets. Furthermore, additional data from questionnaires, class attendance or data from instructors could enhance the performance of this model.

Keyword : Machine Learning, Student Performance, Prediction, Education

กิตติกรรมประกาศ

สารนิพนธ์นี้สำเร็จลุล่วงได้ด้วยความอนุเคราะห์จาก ผศ.ดร.จันตรี ผลประเสริฐ อาจารย์ที่ปรึกษาที่ให้คำปรึกษา คำแนะนำในการทำสารนิพนธ์ ตลอดจนสนับสนุนข้อมูลทางวิชาการและข้อมูลสำหรับทำสารนิพนธ์นี้

ขอกราบขอบพระคุณโรงเรียนมัธยมศึกษาจากจังหวัดสุพรรณบุรี ที่ให้ความอนุเคราะห์ข้อมูลผลการเรียนของนักเรียนและข้อมูลคะแนนรายวิชาของนักเรียนมาใช้ในการทำสารนิพนธ์นี้

ขอกราบขอพระคุณคณะกรรมการสอบสารนิพนธ์ที่ได้ให้คำแนะนำและข้อเสนอแนะสำหรับการปรับปรุงสารนิพนธ์ และการใช้เทคนิคการเรียนรู้ของเครื่องโดยใช้แบบจำลอง Random Forest, Logistic Regression, Support Vector Machine, K-Nearest Neighbor และ XGBoost เพื่อเป็นเครื่องมือในการวิเคราะห์และแก้ไขปัญหาทางข้อมูลต่อไป

ขวัญตา ศิลปไพบูลย์พานิช

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

ภูมิหลัง

การศึกษาของไทยในปัจจุบันนั้นมุ่งเน้นให้นักเรียนได้สำรวจความถนัดของตนเอง เพื่อสามารถกำหนดทิศทางการประกอบอาชีพได้ตรงความต้องการที่สุด การศึกษาจึงมีความสำคัญอย่างมากในทุกระดับชั้นซึ่งจะเป็นพื้นฐานในการศึกษาต่อและการประกอบอาชีพได้เป็นอย่างดี

สำหรับนักเรียนชั้นมัธยมศึกษาตอนต้นนั้นการศึกษาต่อในระดับชั้นมัธยมศึกษาตอนปลายถือเป็นเรื่องสำคัญอย่างหนึ่ง เนื่องจากเป็นการกำหนดทิศทางพื้นฐานในการเลือกสาขาวิชาสำหรับอาชีพในอนาคต ซึ่งสาขาวิชาวิศวกรรมศาสตร์หรือทางวิทยาศาสตร์และเทคโนโลยีถือเป็นหนึ่งในสาขาวิชาที่ได้รับความสนใจจากนักเรียนอย่างมากสาขาหนึ่ง โดยนักเรียนจำนวนมากประสบปัญหาการเลือกสาขาวิชาเรียน ซึ่งสาเหตุที่พบมากคือ นักเรียนไม่สามารถเลือกสาขาวิชาที่สนใจได้เนื่องจากมีผลการเรียนไม่เพียงพอต่อความต้องการของหลักสูตร ซึ่งสถานศึกษานั้นได้กำหนดเกณฑ์การรับเข้าเพื่อศึกษาต่อระดับชั้นมัธยมศึกษาตอนปลายสำหรับแผนการเรียนที่แตกต่างกัน สำหรับเกณฑ์การสมัครเข้าศึกษาต่อในสาขาวิชาวิศวกรรมศาสตร์หรือทางวิทยาศาสตร์และเทคโนโลยีนั้นโดยมากพิจารณาจากรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษเป็นสำคัญ ในแต่ละปีการศึกษามีนักเรียนจำนวนไม่น้อยที่ประสบปัญหาดังกล่าวเนื่องจากไม่สามารถปรับปรุงแก้ไขผลการเรียนของตนเองได้ทันก่อนจบการศึกษาระดับชั้นมัธยมศึกษาตอนต้น

จากเหตุผลดังกล่าวทำให้ผู้วิจัยเล็งเห็นถึงความสำคัญของผลการเรียนที่ส่งผลกระทบต่อการเลือกสาขาวิชาวิศวกรรมศาสตร์หรือทางวิทยาศาสตร์และเทคโนโลยีในระดับชั้นมัธยมศึกษาตอนปลาย หากสามารถประมาณการผลการเรียนของนักเรียนได้จะมีประโยชน์อย่างมากในการช่วยเหลือให้นักเรียนประสบความสำเร็จในการเข้าศึกษาต่อในสาขาวิชาที่สนใจ นอกจากนี้ครูผู้สอนจะสามารถให้คำแนะนำและช่วยเหลือนักเรียนได้ทันเวลา

เนื่องจากจำนวนนักเรียนที่จะจบการศึกษาในระดับชั้นมัธยมศึกษาตอนต้นมีจำนวนค่อนข้างมาก ซึ่งหากใช้การประมาณการด้วยครูหรือบุคลากรนั้นอาจทำให้การดำเนินการล่าช้าจากข้อจำกัดด้านจำนวนของครูและบุคลกร ผู้วิจัยจึงเล็งเห็นถึงความสำคัญที่จะใช้เทคโนโลยีการประมาณการเข้ามาช่วยเหลือ ซึ่งปัจจุบันมีเทคโนโลยีที่สามารถช่วยแก้ไขปัญหาดังกล่าวมากมาย หนึ่งในเทคนิคนั้นคือ การเรียนรู้ของเครื่องที่สามารถสร้างแบบจำลองการเรียนรู้ข้อมูลและทำนายผลได้โดยอาศัยชุดข้อมูลของนักเรียน เช่น คะแนนรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และ

ภาษาอังกฤษในอดีต ซึ่งอาจเป็นประโยชน์ในงานด้านการทำนายผลการเรียนของนักเรียนที่จะช่วยปรับปรุงผลการเรียนให้ดีขึ้นก่อนจบการศึกษา

งานวิจัยนี้มีจุดประสงค์เพื่อศึกษาการนำเทคนิคการเรียนรู้ของเครื่องเพื่อใช้ในการทำนายผลการเรียนรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษ ซึ่งในการทำนายจะใช้เทคนิค Extreme Gradient Boosting (XGBoost) , Logistic Regression , Support Vector Machine (SVM) , K Nearest Neighbor (KNN) และ Random Forest เพื่อเปรียบเทียบความแม่นยำและสามารถนำแบบจำลองนั้นมาช่วยปรับปรุงการเรียนการสอนให้มีประสิทธิภาพมากยิ่งขึ้น

ความมุ่งหมายของการวิจัย

ในการวิจัยครั้งนี้ผู้วิจัยได้ตั้งความมุ่งหมายไว้ดังนี้

1. ศึกษาการนำเทคนิคการเรียนรู้ของเครื่องเพื่อใช้ในการทำนายผลการเรียนรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษ โดยใช้คะแนนรายวิชาในอดีต
2. ศึกษาเทคนิค Feature Engineering เพื่อหา feature ที่ใช้ในการทำนายได้อย่างแม่นยำ
3. ศึกษาเทคนิค Synthetic Minority Oversampling Technique (SMOTE) เพื่อช่วยในการจัดการข้อมูลที่ไม่สมดุล (Imbalance dataset)

ความสำคัญของการวิจัย

งานวิจัยนี้ศึกษาการทำนายผลการเรียนของนักเรียนล่วงหน้า เพื่อประมาณการผลการเรียนของนักเรียนรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษ โดยได้ทำการขอความอนุเคราะห์ข้อมูลที่จัดเก็บจากโรงเรียนระดับมัธยมศึกษาแห่งหนึ่งในจังหวัดสุพรรณบุรีมาใช้ในการศึกษา โดยหากกล่าวถึงเทคนิคการทำนายผลในรูปแบบการจำแนกกลุ่มข้อมูล(Classification) นั้นมี อัลกอริทึมที่ได้รับความนิยมสำหรับการทำนายผล ดังนี้ XGBoost , Logistic Regression , SVM , KNN และ Random Forest ซึ่งผู้วิจัยมีความสนใจและต้องการศึกษาแบบจำลองดังกล่าวเพื่อเปรียบเทียบประสิทธิภาพการทำงาน โดยใช้ Accuracy, Precision, Recall, Confusion matrix และ Macro F1-Score ในการเปรียบเทียบสมรรถนะ

ขอบเขตของการวิจัย

ประชากรที่ใช้ในการวิจัย

ใช้ข้อมูลที่รวบรวมจากโรงเรียนมัธยมศึกษาแห่งหนึ่งในจังหวัดสุพรรณบุรี ซึ่งเป็นข้อมูลผลการเรียนของนักเรียนระดับชั้นมัธยมศึกษาตอนต้นตั้งแต่ปีการศึกษา 2558 - 2560 ประกอบด้วย ข้อมูลคะแนนรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษ

กลุ่มตัวอย่างที่ใช้ในการวิจัย

ใช้ข้อมูลที่รวบรวมจากโรงเรียนมัธยมศึกษาแห่งหนึ่งในจังหวัดสุพรรณบุรี ซึ่งเป็นข้อมูลผลการเรียนรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษของนักเรียนระดับชั้นมัธยมศึกษาตอนต้นตั้งแต่ปีการศึกษา 2558 - 2560 จำนวน 1,382 คน และมีข้อมูลคุณลักษณะทั้งหมด 15 คอลัมน์

ตัวแปรที่ศึกษา

1. ตัวแปรอิสระ แบ่งเป็นดังนี้

1.1 ข้อมูลคะแนนรายวิชาคณิตศาสตร์

1.1.1 Math1 (คะแนนรายวิชาคณิตศาสตร์ 1)

1.1.2 Math2 (คะแนนรายวิชาคณิตศาสตร์ 2)

1.1.3 Math3 (คะแนนรายวิชาคณิตศาสตร์ 3)

1.1.4 Math4 (คะแนนรายวิชาคณิตศาสตร์ 4)

1.1.5 Math5 (คะแนนรายวิชาคณิตศาสตร์ 5)

1.2 ข้อมูลคะแนนรายวิชาวิทยาศาสตร์

1.2.1 Sci1 (คะแนนรายวิชาวิทยาศาสตร์ 1)

1.2.2 Sci2 (คะแนนรายวิชาวิทยาศาสตร์ 2)

1.2.3 Sci3 (คะแนนรายวิชาวิทยาศาสตร์ 3)

1.2.4 Sci4 (คะแนนรายวิชาวิทยาศาสตร์ 4)

1.2.5 Sci5 (คะแนนรายวิชาวิทยาศาสตร์ 5)

1.3 ข้อมูลคะแนนรายวิชาภาษาอังกฤษ

1.3.1 Eng1 (คะแนนรายวิชาภาษาอังกฤษ 1)

1.3.2 Eng2 (คะแนนรายวิชาภาษาอังกฤษ 2)

1.3.3 Eng3 (คะแนนรายวิชาภาษาอังกฤษ 3)

1.3.4 Eng4 (คะแนนรายวิชาภาษาอังกฤษ 4)

1.3.5 Eng5 (คะแนนรายวิชาภาษาอังกฤษ 5)

2. ตัวแปรตาม ได้แก่ ผลการเรียนรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษ

กรอบแนวคิดในงานวิจัย

งานวิจัยนี้พัฒนาแบบจำลองการทำนายโดยใช้เทคนิคการเรียนรู้ของเครื่องที่ใช้การสั่งงานด้วยระบบคอมพิวเตอร์ โดยใช้ข้อมูลที่ได้รวบรวมเรียบร้อยแล้วจากโรงเรียนมัธยมศึกษาแห่งหนึ่งในจังหวัดสุพรรณบุรี ซึ่งข้อมูลผลการเรียนที่ใช้ในการทำนาย ประกอบด้วยคุณลักษณะทั้งหมด 15 คอลัมน์ ได้แก่ Math1 (คะแนนรายวิชาคณิตศาสตร์ 1) , Math2 (คะแนนรายวิชาคณิตศาสตร์ 2) , Math3 (คะแนนรายวิชาคณิตศาสตร์ 3) , Math4 (คะแนนรายวิชาคณิตศาสตร์ 4) , Math5 (คะแนนรายวิชาคณิตศาสตร์ 5) , Sci1 (คะแนนรายวิชาวิทยาศาสตร์ 1) , Sci2 (คะแนนรายวิชาวิทยาศาสตร์ 2) , Sci3 (คะแนนรายวิชาวิทยาศาสตร์ 3) , Sci4 (คะแนนรายวิชาวิทยาศาสตร์ 4) , Sci5 (คะแนนรายวิชาวิทยาศาสตร์ 5) , Eng1 (คะแนนรายวิชาภาษาอังกฤษ 1) , Eng2 (คะแนนรายวิชาภาษาอังกฤษ 2) , Eng3 (คะแนนรายวิชาภาษาอังกฤษ 3) , Eng4 (คะแนนรายวิชาภาษาอังกฤษ 4) และEng5 (คะแนนรายวิชาภาษาอังกฤษ 5) โดยมีข้อมูลตั้งแต่ปีการศึกษา 2558 - 2560 ในการทำนายผลการเรียนรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษ

สมมุติฐานในการวิจัย

1. แบบจำลอง XGBoost, แบบจำลอง Logistic regression, แบบจำลอง SVM, แบบจำลอง KNN และแบบจำลอง Random Forest สามารถทำนายผลการเรียนของนักเรียนล่วงหน้าได้

2. แบบจำลอง XGBoost มีประสิทธิภาพการทำนายดีที่สุด

3. การใช้เทคนิค Feature Engineering จะช่วยเพิ่มประสิทธิภาพการทำนายได้อย่างแม่นยำยิ่งขึ้น

4. การใช้เทคนิค SMOTE จะช่วยจัดการปัญหาข้อมูลที่ไม่สมดุลได้

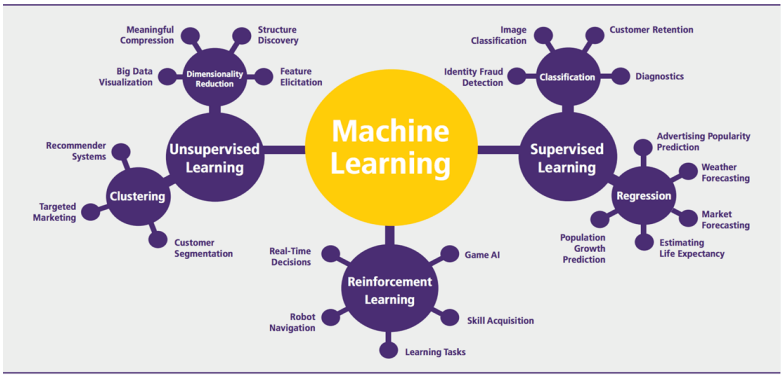
บทที่ 2
ทบทวนวรรณกรรม

ในการวิจัยครั้งนี้ ผู้วิจัยได้ทำการศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง และได้นำเสนอตามหัวข้อต่อไปนี้

1. การเรียนรู้ของเครื่อง (Machine Learning)
2. การเพิ่มประสิทธิภาพของแบบจำลองการทำนายด้วยเทคนิค Feature Engineering
3.การคัดเลือกคุณลักษณะสำหรับแบบจำลองการทำนายด้วยเทคนิค Feature Selection
4. การจัดการกับข้อมูลที่ไม่สมดุล
5. เทคนิคการจำแนกกลุ่มข้อมูล
6. การวัดประสิทธิภาพของอัลกอริทึม
7. งานวิจัยที่เกี่ยวข้อง

การเรียนรู้ของเครื่อง (Machine Learning)

การเรียนรู้ของเครื่องเป็นศาสตร์ทางด้านวิทยาศาสตร์ที่เกี่ยวข้องกับอัลกอริทึมและแบบจำลองทางสถิติ ซึ่งใช้การสั่งงานโดยระบบคอมพิวเตอร์ในรูปแบบเชิงการทำนายผลจากข้อมูลในการทำนายผลอาศัยข้อมูลส่วนหนึ่งในการเรียนรู้ เรียกว่า ชุดข้อมูลการฝึกอบรม (Training Data) เพื่อให้แบบจำลองได้ฝึกฝนและเรียนรู้ลักษณะของข้อมูลและคำตอบ จากนั้นจึงทำการทดสอบแบบจำลองด้วยข้อมูลอีกส่วนหนึ่ง เรียกว่า ชุดข้อมูลทดสอบ(Testing Data) หลักการทำงานของการเรียนรู้ของเครื่องนั้นอาศัยการคำนวณทางสถิติ ซึ่งมุ่งเน้นไปที่การทำนายโดยใช้คอมพิวเตอร์เป็นหลัก (วิกิพีเดีย, 2562) ปัจจุบันมีการแบ่งประเภทของการเรียนรู้ของเครื่องเป็น 3 ประเภท ได้แก่ การเรียนรู้แบบมีผู้สอน(Supervised Learning) , การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) และ การเรียนรู้แบบเสริมกำลัง(Reinforcement Learning) ดังแสดงในรูปภาพที่ 1



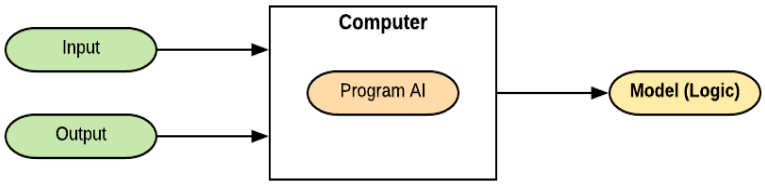
ภาพประกอบ 1 แสดงการแบ่งประเภทของการเรียนรู้ของเครื่อง

ที่มา: Affan Samad. (2019). What is Machine Learning. Retrieved from https://becominghuman.ai/what-is-machine-learning-d292114cc6ce

การเรียนรู้แบบมีผู้สอน(Supervised Learning)

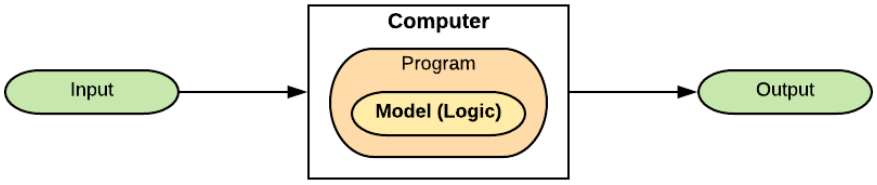
เทคนิคการเรียนรู้ของเครื่องประเภทการเรียนรู้แบบมีผู้สอน มีหลักการเรียนรู้จากข้อมูลนำเข้า(Input Data)ที่จัดเตรียมไว้ก่อนหน้า ซึ่งประกอบด้วยข้อมูลคุณลักษณะ(Feature)และข้อมูลคำตอบ(Label) จากนั้นจึงทำการทำนายผลลัพธ์(Output Data)จากสิ่งที่ได้เรียนรู้ ด้วยลักษณะการทำนายนี้ทำให้ในขั้นตอนการเรียนรู้ต้องอาศัยข้อมูลจำนวนค่อนข้างมากเพื่อให้ได้ผลการทำนายที่แม่นยำมากขึ้น เทคนิคการเรียนรู้แบบมีผู้สอนที่นิยม ได้แก่ การจำแนกประเภท (Classification) และการถดถอย(Regression)(วิกิพีเดีย, 2557)

โดยข้อมูลที่ใช้สำหรับการเรียนรู้จะต้องมีคุณสมบัติที่สนับสนุนการเรียนรู้และการแก้ปัญหาเรียกว่า Feature Vector ในขั้นตอนการเรียนรู้ของเครื่อง(Training)นั้นจะดำเนินการโดยใช้อัลกอริทึม(Feature)ที่เหมาะสมกับลักษณะของปัญหาที่ต้องการแก้ไข จนกระทั่งได้วิธีการที่ดีที่สุดสำหรับสร้างเป็นแบบจำลองดังแสดงในรูปภาพที่ 2 หลังจากทำขั้นตอนการเรียนรู้จะนำแบบจำลองที่ได้มาประยุกต์ใช้งาน โดยการนำชุดข้อมูลอีกหนึ่งชุด เรียกว่า Test data เข้าสู่แบบจำลอง เพื่อทำนาย(Prediction)ผลลัพธ์ ดังแสดงในรูปภาพที่ 3(Chalermkiatsakul, 2018)



ภาพประกอบ 2 แสดงกระบวนการเรียนรู้ของเครื่องเพื่อให้ได้มาซึ่งแบบจำลอง

ที่มา: Phuri Chalermkiatsakul. (2018). Supervised Learning. Retrieved from https://medium.com/@every.phu/supervised-learning-คืออะไร-ทำงานยังไง-1c0e411a40a2



ภาพประกอบ 3 แสดงกระบวนการประยุกต์ใช้งานแบบจำลองเพื่อทำนายผล

ที่มา: Phuri Chalermkiatsakul. (2018). Supervised Learning. Retrieved from https://medium.com/@every.phu/supervised-learning-คืออะไร-ทำงานยังไง-1c0e411a40a2

การเรียนรู้แบบไม่มีผู้สอน(Unsupervised Learning)

เทคนิคการเรียนรู้ของเครื่องประเภทการเรียนรู้แบบไม่มีผู้สอน มีหลักการเรียนรู้จากข้อมูลนำเข้าที่มีเพียงข้อมูลคุณลักษณะเท่านั้น โดยการเรียนรู้และการทำนายผลนั้นจะอาศัยการจำแนกและสร้างรูปแบบจากข้อมูลนำเข้าโดยจะไม่มีการกำหนดผลลัพธ์ที่ได้ล่วงหน้า เทคนิคที่นิยม ได้แก่ การลดมิติของข้อมูล (Dimensionality Reduction) และการจัดกลุ่ม(Clustering) (วิกิพีเดีย, 2556)

การเรียนรู้แบบเสริมกำลัง(Reinforcement Learning)

เทคนิคการเรียนรู้แบบเสริมกำลังมีหลักการทำงานในลักษณะของการให้รางวัล(Reward)หรือลงโทษ ซึ่งจะมีการตั้งเป้าความสำเร็จไว้ก่อนล่วงหน้า โดยในแต่ละสถานการณ์จะมีตัวเลือกของการกระทำ(Action)ให้เลือก ซึ่งแบบจำลองจะเรียนรู้จากสภาพแวดล้อม

(Environment)นั้นๆ เพื่อหาตัวเลือกการกระทำที่ดีที่สุดในแต่ละสถานการณ์ที่พบ (วิกิพีเดีย, 2556) เป้าหมายของการกระทำคือการเรียนรู้เพื่อหาวิธีการให้ได้รับรางวัลจากสถานการณ์ต่างๆรวมกันให้ได้มากที่สุด(Maximum Sum of Expected Rewards) (Vijite, 2018)

การเพิ่มประสิทธิภาพของแบบจำลองการทำนายด้วยเทคนิค Feature Engineering

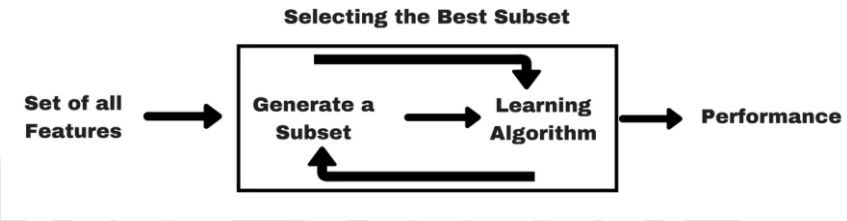
Feature Engineering คือการแปลงข้อมูลดิบ(Raw Data) ในทางคณิตศาสตร์เพื่อสร้างคุณลักษณะใหม่ขึ้นมาเพื่อปรับปรุงความแม่นยำของแบบจำลองการเรียนรู้ของเครื่องให้ดียิ่งขึ้น ทั้งในด้านของความสามารถในการสร้างคุณลักษณะที่สัมพันธ์กับเป้าหมาย(Label), ความสามารถในการนำแหล่งข้อมูลภายนอก(External Data Sources)มาใช้กับแบบจำลอง, ความสามารถในการใช้งานข้อมูลแบบไม่มีโครงสร้าง(Unstructured Data), ความสามารถในการสร้างคุณลักษณะที่ตีความ(Interpretable)ได้ดีกว่า, ความสามารถในการสร้างคุณลักษณะจำนวนมากได้อย่างอิสระและความสามารถในการเลือกชุดข้อมูลย่อย(Subset)ผ่านขั้นตอนการเลือกคุณลักษณะ(Feature Selection) เป็นต้น (Odegua, 2015)

Feature Engineering มีลักษณะการเรียนรู้แบบDomain Knowledge สำหรับสร้าง Featureใหม่ขึ้นและคัดเลือกคุณลักษณะที่ไม่สัมพันธ์กับข้อมูลคำตอบออก เพื่อให้แบบจำลองได้เรียนรู้จากข้อมูลที่มีความสำคัญที่สุด ส่งผลให้การทำนายมีผลลัพธ์ที่ดีขึ้น ซึ่งมีวิธีการที่ใช้ในขั้นตอนการทำFeature Engineeringหลากหลาย อาทิเช่น การเพิ่มข้อมูลในส่วนที่ขาดหายไป (Imputation), การจัดการกับข้อมูลรบกวน(Noisy Data), การแปลงข้อมูลประเภทหมวดหมู่ (Category)ให้เป็นตัวเลขที่มีความสำคัญเท่ากัน, การจัดการข้อมูลเพื่อให้มีค่าอยู่ในช่วงเดียวกัน (Scaling) เป็นต้น (Wikipedia, 2019)

สำหรับงานวิจัยนี้ใช้เทคนิคการทำ Polynomial Features เพื่อช่วยเพิ่มประสิทธิภาพการทำงานของแบบจำลองให้ดียิ่งขึ้น โดยทั่วไปมักใช้กับแบบจำลองที่มีจำนวนข้อมูลคุณลักษณะน้อย หรือใช้ในกรณีที่ประสิทธิภาพของคุณลักษณะหนึ่งขึ้นอยู่กับคุณลักษณะหนึ่ง เทคนิค Polynomial Features จะช่วยเพิ่มความซับซ้อนให้กับข้อมูลและเพิ่มความยืดหยุ่นให้กับแบบจำลอง โดยการสร้างFeature Matrixใหม่จากการรวมคุณลักษณะหลายๆคุณลักษณะด้วยการคูณค่ายกกำลัง โดยจำนวนของFeature Matrixที่ได้ขึ้นอยู่กับการกำหนดพารามิเตอร์ค่ายกกำลังสูงสุดของคุณลักษณะ (degree) (Dorpe, 2018) นอกจากนี้จะใช้เทคนิค Scaling ในการจัดการข้อมูลคุณลักษณะให้มีค่าอยู่ในช่วงเดียวกัน

การคัดเลือกคุณลักษณะสำหรับแบบจำลองการทำนาย(Feature Selection)

Feature Selection คือขั้นตอนสำหรับคัดเลือกคุณลักษณะที่มีความสำคัญกับการสร้างแบบจำลองให้มีคุณภาพ นอกจากนี้ยังมีประโยชน์ในด้านการลดจำนวนคุณลักษณะของชุดข้อมูลที่มีจำนวนมากเกินความจำเป็น เพื่อลดระยะเวลาในการสร้างแบบจำลองโดยจะคัดเลือกคุณลักษณะที่มีความสัมพันธ์(Correlation)สูงกับข้อมูลคำตอบ ซึ่งจะส่งผลให้ประสิทธิภาพในการทำงานของแบบจำลองดียิ่งขึ้น ในขณะเดียวกันคุณลักษณะนั้นต้องมีความสัมพันธ์ระหว่างคุณลักษณะด้วยกันต่ำ เพื่อให้ข้อมูลมีความเป็นอิสระต่อกันมากที่สุด สำหรับเทคนิค Feature Selection สามารถแบ่งได้ 3 วิธีหลักๆ คือ วิธีฟิลเตอร์(Filter Methods) , วิธีแรปเปอร์(Wrapper Methods) และ วิธีฝังตัว(Embedded Methods) (arnondora, 2019)



ภาพประกอบ 4 แสดงกระบวนการคัดเลือกคุณลักษณะแบบวิธีแรปเปอร์

ที่มา: Saurav Kaushik. (2016). Introduction to Feature Selection methods with an example. Retrieved from https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/

สำหรับงานวิจัยนี้จะใช้วิธีการคัดเลือกคุณลักษณะแบบวิธีแรปเปอร์ หลักการทำงานคือการสร้างSubsetของคุณลักษณะ จากนั้นจะทำการหาค่าความแม่นยำของการจำแนกกลุ่มข้อมูลในชุดข้อมูลการเรียนรู้ หรืออีกกรณีหนึ่งคือใช้ค่าความแม่นยำของการจำแนกกลุ่มข้อมูลในการวัดความสำคัญของSubset ซึ่งหนึ่งในวิธีการคัดเลือกคุณลักษณะแบบแรปเปอร์ที่นิยม คือ Recursive Feature Elimination (RFE) การคัดเลือกคุณลักษณะด้วย RFE เหมาะสมกับปัญหาการจำแนกที่มีข้อมูลกลุ่มตัวอย่าง(sample)จำนวนน้อย ซึ่งเดิมทีถูกนำไปใช้กับการจำแนกประเภทข้อมูลไมโครอาร์เรย์(microarray)ในโรคมะเร็ง ซึ่งมีข้อมูลตัวอย่างของชุดข้อมูลการเรียนรู้น้อยกว่า 100 เรคคอร์ด ในขณะที่จำนวนของคุณลักษณะนั้นมีจำนวนหลายพันเรคคอร์ด ส่งผลให้ RFE กลายเป็นแนวทางที่มีประสิทธิภาพในการคัดเลือกคุณลักษณะของกลุ่มตัวอย่างขนาดเล็ก (Chen และ Jeong, 2008)

การคัดเลือกคุณลักษณะด้วยวิธีการแบบ RFE นั้นจะใช้กับชุดข้อมูลการเรียนรู้เพื่อจุดประสงค์ในการคัดเลือกคุณลักษณะหรือตัวทำนาย(predictor) รวมถึงเพื่อหาจำนวนคุณลักษณะที่เหมาะสมสำหรับใช้ในการทำนาย หรือใช้เพื่อปรับแต่งและประเมินประสิทธิภาพของแบบจำลองนั้นๆ ซึ่งจะทำการประเมินจากค่าสัมประสิทธิ์เพื่อคัดคุณลักษณะออกไปตามลำดับ โดยพิจารณจากคุณลักษณะที่มีความสำคัญน้อยที่สุด, คุณลักษณะที่มีค่าความน่าจะเป็นสูงที่สุด หรือมีค่าความน่าจะเป็นสูงกว่าระดับนัยสำคัญและคุณลักษณะที่ซ้ำซ้อนหรือมีความสำคัญใกล้เคียงกัน เป็นต้น ซึ่งการกระทำดังกล่าวจะทำวนซ้ำไปจนกว่าจำนวนคุณลักษณะที่ได้นั้นจะเท่ากับจำนวนคุณลักษณะที่ได้ทำการกำหนดไว้ (developers, 1999) การทำงานของ RFE ถือได้ว่ามีการจัดอันดับเพื่อคัดคุณลักษณะออกที่ค่อนข้างดี นำไปสู่การลดข้อผิดพลาดการทำนายผลลัพธ์ในชุดข้อมูลเดียวกัน

การจัดการกับข้อมูลที่ไม่สมดุล (Imbalance Dataset)

ปัญหา Imbalance Dataset สำหรับการเรียนรู้ของเครื่องนั้นเป็นปัญหาที่พบได้ในกรณีของเทคนิคการจำแนกกลุ่มข้อมูล กล่าวคือปัญหาเกิดจากการมีจำนวนข้อมูลในแต่ละกลุ่ม(Class)ที่ไม่เท่ากัน ซึ่งเป็นหนึ่งในปัจจัยที่มีผลต่อประสิทธิภาพการทำงานของแบบจำลอง หากข้อมูลในแต่ละClassมีจำนวนที่ใกล้เคียงกันจะทำให้กระบวนการเรียนรู้ของแบบจำลองมีประสิทธิภาพที่ดีขึ้น แต่หากข้อมูลในแต่ละClassมีจำนวนที่ต่างกันมากอาจทำให้เกิดปัญหาความเหลื่อมล้ำ(Bias) ซึ่งจะส่งผลให้ได้ค่าความแม่นยำสูงในการทำนายClassที่มีข้อมูลจำนวนมากและได้ค่าความแม่นยำต่ำในClassที่มีจำนวนข้อมูลน้อย ทำให้ผลการทำนายโดยรวมไม่มีประสิทธิภาพเท่าที่ควร (Brink, Richards, และ Fetherolf, 2016)

งานวิจัยนี้ใช้วิธีการทำ Sampling Methods ด้วยเทคนิค SMOTE สำหรับการลดปัญหา Imbalance Dataset ซึ่งเป็นเทคนิคที่มีประสิทธิภาพและได้รับความนิยมอย่างมาก มีหลักการทำงานโดยการสุ่มตัวอย่างข้อมูล(Data Point) จากClassที่มีข้อมูลจำนวนน้อย(Minority Class) จากชุดข้อมูลเดิม เพื่อเพิ่มจำนวนข้อมูลของClassนั้นๆให้มีจำนวนเท่ากับClassที่มีจำนวนข้อมูลมากที่สุดโดยเท่าๆกันทุกคลาส (Jordan, 2018) รูปภาพที่ 5 แสดงหลักการทำงานของเทคนิค SMOTE จากการสังเคราะห์ข้อมูลขึ้นใหม่จากข้อมูลเดิม โดยการประมาณค่าจากตำแหน่งที่มีระยะห่างใกล้เพื่อนบ้านที่สุด (K-Nearest Neighbor)

$$\mathbf{x_{new}} = \mathbf{x_i} + \lambda(\mathbf{x_{zi}} - \mathbf{x_i}) \quad (1)$$

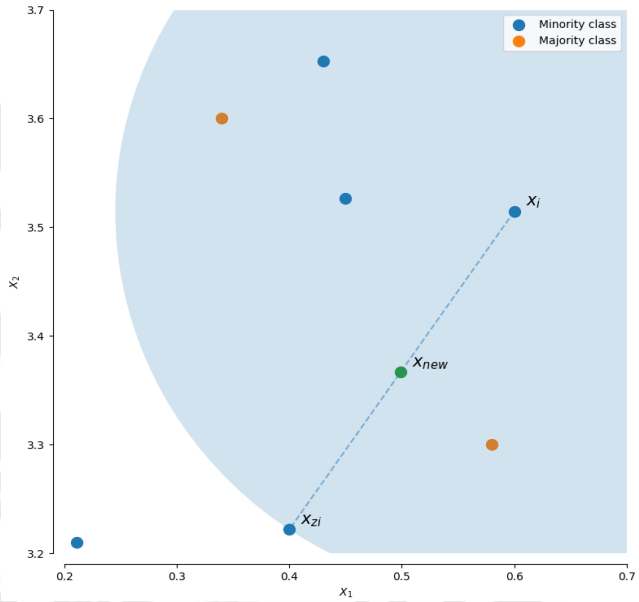
จากสมการ

$\mathbf{x_{new}}$ คือ ผลลัพธ์จากการทำเทคนิค SMOTE

$\mathbf{x_i}$ คือ ข้อมูลที่สังเคราะห์ขึ้นใหม่

$\mathbf{x_{zi}}$ คือ ข้อมูลเพื่อนบ้านที่ใกล้ที่สุด(K-Nearest Neighbor)

λ คือ จำนวนการสุ่มตัวอย่างในช่วง 0-1 ระหว่าง $\mathbf{x_i}$ และ $\mathbf{x_{zi}}$



ภาพประกอบ 5 แสดงการเพิ่มจำนวนข้อมูลของเทคนิค SMOTE

ที่มา: Jeremy Jordan. (2018). Learning From Imbalanced Data. Retrieved from https://www.jeremyjordan.me/imbalanced-data/

เทคนิคการจำแนกกลุ่มข้อมูล (Classification)

Classification เป็นเทคนิคหนึ่งของการเรียนรู้ของเครื่องแบบมีผู้สอน โดยกระบวนการเรียนรู้นั้นจำเป็นต้องมีข้อมูลคำตอบ ซึ่งจะแบ่งข้อมูลคำตอบออกเป็นClassต่างๆ สำหรับหลักการทำงานของแบบจำลองจะแบ่งข้อมูลออกเป็น 2 ส่วน คือชุดข้อมูลการเรียนรู้และชุดข้อมูลทดสอบ ซึ่งในงานวิจัยนี้แบ่งข้อมูลดังกล่าวในอัตราส่วน 80 : 20 หรือตามความเหมาะสม จากนั้นนำข้อมูลชุดการเรียนรู้ป้อนเข้าสู่ระบบเพื่อเข้าสู่ขั้นตอนการเรียนรู้โดยใช้อัลกอริทึมที่กำหนด จนกระทั่งได้ผลลัพธ์ออกมาเป็นแบบจำลองการทำนาย จากนั้นนำข้อมูลชุดทดสอบซึ่งเป็นข้อมูลที่เตรียมไว้

สำหรับทดสอบประสิทธิภาพการทำงานของแบบจำลองมาทำใช้ทำนายผลและพิจารณาผลลัพธ์ที่ได้ โดยในงานวิจัยนี้ได้ใช้เทคนิคในการจำแนกกลุ่มข้อมูล ดังนี้

ทฤษฎีอัลกอริทึม Logistic Regression

การวิเคราะห์การถดถอยโลจิสติก(Logistic Regression)เป็นเทคนิคที่ใช้ในการวิเคราะห์ข้อมูลเชิงคุณภาพ(Qualitative data) ซึ่งลักษณะของข้อมูลนั้นจะเป็นประเภทหมวดหมู่ การวิเคราะห์การถดถอยโลจิสติกมักถูกใช้กับงานด้านการจำแนกกลุ่มข้อมูล เพื่อทำนายความน่าจะเป็นของการเกิดของเหตุการณ์ต่างๆ โดยสามารถแบ่งประเภทการวิเคราะห์ได้ 3 ประเภทหลักๆ ดังนี้ (1.)การวิเคราะห์การถดถอยโลจิสติกแบบไบนารี(Binary Logistic Regression Analysis) คือการวิเคราะห์ข้อมูลโดยที่ค่าของคำตอบ(Labels) มีความน่าจะเป็นเพียง 2 กลุ่ม (2.)การวิเคราะห์การถดถอยโลจิสติกพหุกลุ่ม(Multinomial Logistic Regression Analysis) เป็นการวิเคราะห์ข้อมูลโดยที่ค่าของคำตอบมีความน่าจะเป็นมากกว่า 2 กลุ่ม ซึ่งเป็นค่าที่ไม่สามารถจัดลำดับได้ (3.)การวิเคราะห์การถดถอยโลจิสติกเชิงลำดับ(Ordinal Logistic Regression) เป็นการวิเคราะห์ข้อมูลโดยที่ค่าของคำตอบมีความน่าจะเป็นมากกว่า 2 กลุ่ม ซึ่งเป็นค่าที่สามารถจัดลำดับได้ โดยหลักการทำงานจะต้องมีการกำหนดขอบเขตการตัดสินใจ(Decision Boundaries) เพื่อระบุว่าข้อมูลที่ทำนายควรจะอยู่กลุ่มใด จากนั้นจะทำการประมาณค่าสัมประสิทธิ์การถดถอยด้วยเทคนิคต่างๆ โดยเทคนิคที่นิยมได้แก่ Maximum Likelihood เป็นต้น(Chandrayan, 2015)

สำหรับงานวิจัยนี้ใช้การวิเคราะห์การถดถอยโลจิสติกเชิงลำดับ(Ordinal Logistic Regression) เนื่องจากค่าคำตอบของชุดข้อมูลที่ใช้ทำนายผลการเรียนเป็นประเภท Ordinal กล่าวคือระดับผลการเรียนของนักเรียนในรายวิชาคณิตศาสตร์ วิทยาศาสตร์และภาษาอังกฤษ แบ่งเป็น 3 ระดับ คือ ระดับดีมาก ระดับดี ระดับพอใช้ และมีการปรับค่าพารามิเตอร์หลักในการสร้างแบบจำลองทำนาย ได้แก่

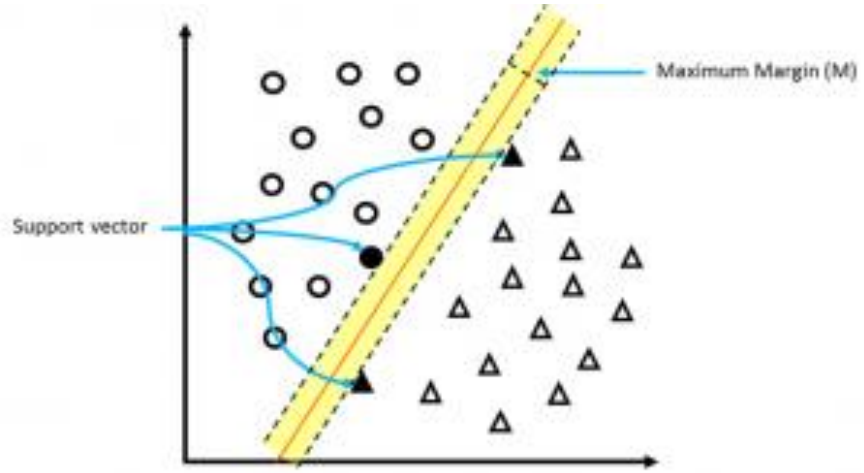
Penalty คือ การระบุ norm ใน penalization

C คือ ค่าผกผันของการทำให้แบบจำลองเป็นมาตรฐาน

ทฤษฎีอัลกอริทึม Support Vector Machine (SVM)

SVM เป็นเทคนิคหนึ่งสำหรับงานด้านการจำแนกกลุ่มข้อมูลที่สามารถทำงานได้อย่างมีประสิทธิภาพ แม้ในกรณีที่ชุดข้อมูลมีจำนวนมิติ(Dimensions)มากกว่าจำนวนตัวอย่าง (Samples) โดยสามารถจัดการทั้งปัญหาเชิงเส้นตรง(Linear Problem) และปัญหาที่ไม่ใช่เชิงเส้นตรง(Non-Linear Problem) สำหรับหลักการทำงานจะเริ่มจากการวางข้อมูลบน Feature Space และสร้างเส้นแบ่งกลุ่มข้อมูล(Hyperplane)เพื่อแบ่งข้อมูลออกเป็นกลุ่มๆ โดยHyperplane ที่ดีนั้นจะต้องมี Marginกว้างที่สุดหรือที่เรียกว่า Maximum Margin ที่สามารถจำแนกข้อมูลแต่ละ

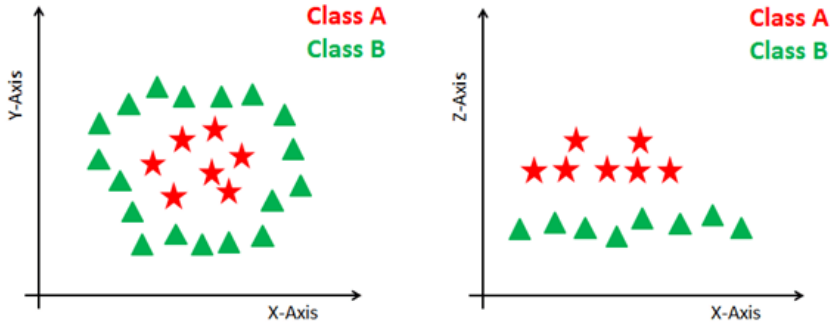
กลุ่มออกจากกันได้อย่างชัดเจน รูปภาพที่ 6 แสดงการจำแนกข้อมูล 2 กลุ่ม โดยที่สมาชิกของข้อมูลแต่ละกลุ่มที่อยู่ใกล้ Hyperplane มากที่สุด เรียกว่า Support Vector มีระยะห่างจาก Hyperplane ที่กว้างเท่าๆกัน อาจเรียกได้ว่ามีขนาดของ Margin ที่กว้างเท่าๆกัน



ภาพประกอบ 6 แสดงการจำแนกข้อมูลด้วยเทคนิค SVM

ที่มา: Avinash Navlani. (2018). Support Vector Machines with Scikit-learn. Retrieved from https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python#svm

อีกหนึ่งความสามารถของ SVM คือการจัดการกับปัญหากรณี Non-Linear Problem คือการจัดการกับข้อมูลที่ไม่สามารถใช้ Linear Hyperplane ได้ ดังแสดงในรูปภาพที่ 7 ซึ่งอธิบายหลักการทำงานของ SVM ภาพด้านซ้ายแสดงภาพของข้อมูลจำนวน 2 กลุ่มที่ไม่สามารถทำการจำแนกได้ด้วยเส้นตรง ภาพทางด้านขวาแสดงการใช้เทคนิค Kernel มาช่วยเพิ่มความสามารถในการจำแนกข้อมูล โดยการแปลงพื้นที่ข้อมูลนำเข้า(Input Space) เป็นพื้นที่ของมิติที่สูงกว่า ซึ่งแต่ละ DataPoints จะถูกPlotลงบนแกน X และแกน Z (โดยที่ Z คือ ผลรวมยกกำลังสองของ X และ Y : Z = X^2=Y^2) (Navlani, 2018)



ภาพประกอบ 7 แสดงการทำงานของ Kernel เพื่อแปลง Input Space
เป็น Higher Dimensional Space

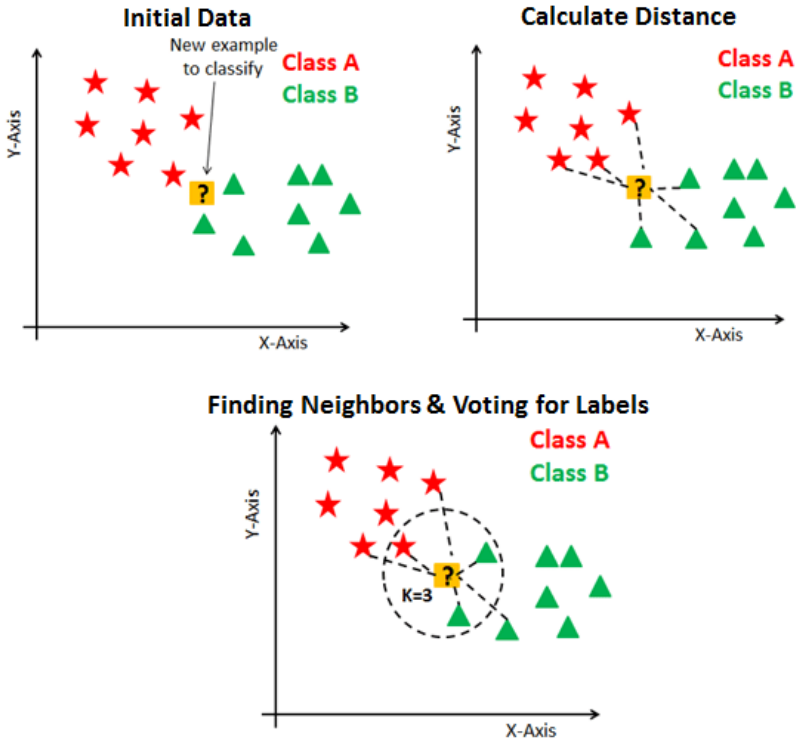
ที่มา: Avinash Navlani. (2018). Support Vector Machines with Scikit-learn. Retrieved from https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python#svm

สำหรับงานวิจัยนี้มีการปรับค่าพารามิเตอร์หลักในการสร้างแบบจำลองทำนาย ได้แก่

Kernel	คือ ประเภทเคอร์เนลที่จะใช้ในอัลกอริทึม
C	คือ Regularization parameter
Gamma	คือ ค่าสัมประสิทธิ์เคอร์เนล

ทฤษฎีอัลกอริทึม K- Nearest Neighbor

K- Nearest Neighbor เป็นเทคนิคที่สามารถใช้จัดการกับปัญหาข้อมูลการถดถอย (Regression Problems)และปัญหาการจำแนกข้อมูล (Classification Problems) ซึ่งเป็นเทคนิคที่ได้รับความนิยมมากเทคนิคหนึ่ง โครงสร้างของแบบจำลองจะถูกกำหนดโดยชุดข้อมูล ดังรูปภาพที่ 8 ซึ่งอธิบายหลักการทำงานของ K- Nearest Neighbor โดยจะมีการกำหนดค่า k หรือจำนวนของเพื่อนบ้านที่ใกล้ที่สุด(Nearest Neighbors) จากนั้นจะทำการคำนวณหาระยะห่างระหว่าง DataPoint กับ Neighbors จำนวน k ตัว และนำระยะห่างที่ได้นั้นมาจัดลำดับ เพื่อหาว่า DataPoint อยู่ใกล้กับNeighborsใดมากที่สุด จากนั้นจึงทำการโหวตโดยยึดตามเสียงข้างมาก(Majority Vote) เพื่อทำนายว่าDataPointนั้นจัดอยู่ในกลุ่มใด โดยดูจากจำนวนของNeighborsส่วนใหญ่ที่อยู่ใกล้กับ DataPoint ดังกล่าว (Latysheva, 2016)



ภาพประกอบ 8 แสดงการทำงานของ K- Nearest Neighbor ในการจำแนกกลุ่มข้อมูล

ที่มา: Avinash Navlani. (2018). KNN Classification using Scikit-learn. Retrieved from https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn

สำหรับงานวิจัยนี้มีการปรับค่าพารามิเตอร์หลักในการสร้างแบบจำลองทำนาย ได้แก่

k_range คือ จำนวน neighbors ที่ใช้ในการทำนาย

leaf_size คือ จำนวนของ sample ที่ query

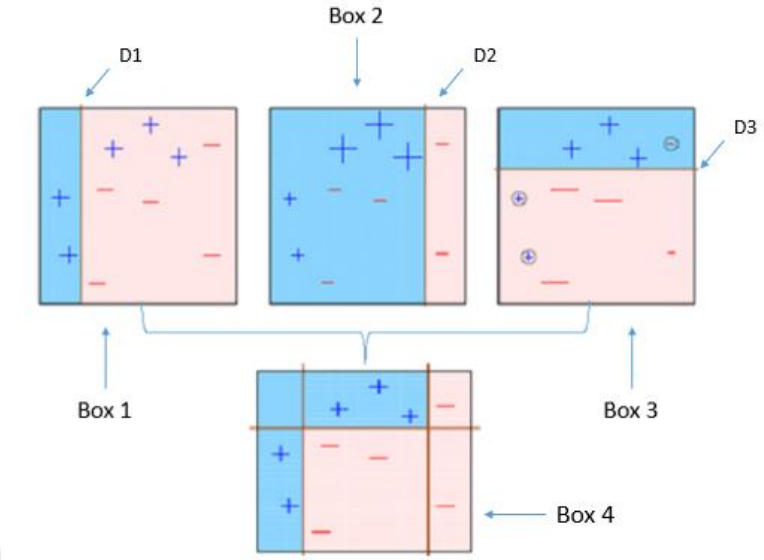
weight_options คือ weight function ที่ใช้ในการทำนาย

metric_option คือ เมตริกที่ใช้วัดระยะใน tree

ทฤษฎีอัลกอริทึม Extreme Gradient Boosting (XGBoost)

XGBoost เป็นเทคนิคที่พัฒนาจากเทคนิค Gradient boosting เพื่อเพิ่มความแม่นยำและความยืดหยุ่นให้กับแบบจำลอง โดยใช้หลักการของ Ensemble Learning Method ในการ Boosting เพื่อสร้างตัวเรียนรู้หลายๆตัว(Multiple Learner)หรือเรียกได้ว่าเป็นการรวม Weak

Learners หลายๆตัวเข้าด้วยกัน ซึ่ง Learner ที่สร้างขึ้นใหม่แต่ละรุ่นนั้นจะทำการแก้ไขข้อบกพร่องในการทำงานของ Learner รุ่นก่อนหน้าเพื่อลด Error ดังแสดงในรูปภาพที่ 9



ภาพประกอบ 9 แสดงการจำแนกกลุ่มข้อมูลด้วยเทคนิค XGBoost

ที่มา: Manish Pathak. (2019). Using XGBoost in Python. Retrieved from https://www.datacamp.com/community/tutorials/xgboost-in-python

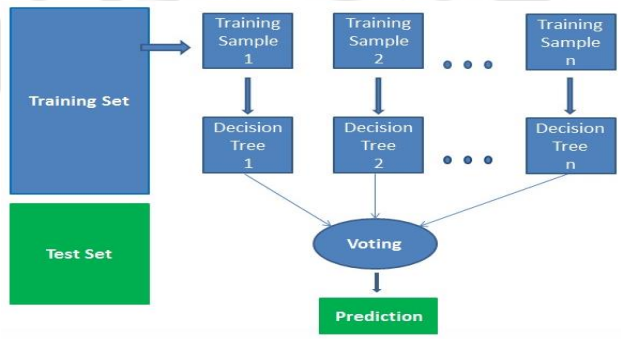
ใน Box 1 มีการสร้างเส้นเพื่อแบ่งข้อมูลครั้งแรกหรือการจำแนกของรุ่นที่ 1 (D1) จากภาพจะแบ่งข้อมูลออกเป็น 2 กลุ่มในแถบสีฟ้าเป็นข้อมูลที่เป็นค่าบวก(+) และแถบสีแดงคือข้อมูลที่เป็นค่าลบ(–) ต่อมาทำการแบ่งข้อมูลรุ่นที่ 2 จะให้ค่าWeightกับการจำแนกข้อมูลที่เป็นค่า + ในแถบสีแดง 3 จุดและสร้างเส้นแบ่งข้อมูล(D2) ผลลัพธ์ได้ตามภาพใน Box 2 ซึ่งจะเห็นได้ว่าการแบ่งข้อมูลในรุ่น 2 นั้นข้อมูลที่เป็นค่า + ทั้ง 3 ค่าที่ทำการแบ่งกลุ่มไปรวมกับค่าที่เป็น – ในกลุ่มแถบสีฟ้า จากนั้นจะทำการแบ่งข้อมูลรุ่นที่ 3 เพื่อแก้ไขข้อผิดพลาดโดยทำการแบ่งข้อมูลใหม่อีกครั้ง โดยการสร้างเส้นแบ่ง (D3) เพื่อแบ่งข้อมูลที่ผิดพลาดในรุ่นที่ 2 ได้ตามภาพ Box 3 จากการจำแนกกลุ่มข้อมูลทั้ง 3 รุ่นนั้นในขั้นตอนสุดท้ายจะทำการรวมค่า Weight ของ Weak Classifiers (Box 1, Box 2 และ Box 3) ซึ่งจะได้ผลลัพธ์ของการทำงานที่ดีขึ้นดังแสดงใน Box 4 (Pathak, 2019)

สำหรับงานวิจัยนี้มีการปรับค่าพารามิเตอร์หลักในการสร้างแบบจำลองทำนาย ได้แก่

n_estimators	คือ weight function ที่ใช้ในการทำนาย
max_depth	คือ เมตริกที่ใช้วัดระยะใน tree
subsample	คือ สัดส่วนของตัวอย่างข้อมูลสำหรับ learner
colsample_bytree	คือ สัดส่วนในการสุ่มคอลัมน์ที่ใช้ในการทำนาย
learning_rate	คือ อัตราควบคุม Weight ในการ train

ทฤษฎีอัลกอริทึม Random Forest

Random Forest เป็นเทคนิคที่พัฒนามาจากเทคนิค Decision Tree สำหรับงานด้านการวิเคราะห์เชิงถดถอยและการจำแนกกลุ่มข้อมูล สำหรับ Random Forest Classificationนั้นจะสร้างแบบจำลองการทำนายในรูปแบบของต้นไม้การตัดสินใจหลายๆต้น(Ensemble of Decision Trees)เพื่อช่วยในการทำนายผลลัพธ์ โดยค่าCorrelationระหว่างDecision Treeแต่ละต้นจะถูกสร้างให้มีความเป็นอิสระต่อกัน(Independent) โดยจะสุ่มเลือกชุดข้อมูลการเรียนรู้แบบ Bootstrap จากชุดข้อมูลการเรียนรู้เดิม ซึ่งชุดข้อมูลที่สุ่มเลือกนั้นมีโอกาสที่จะถูกสุ่มเลือกซ้ำ โดย Decision Treeแต่ละต้นจะสุ่มเลือกชุดข้อมูลคุณลักษณะที่ใช้ในการทำนาย ทำให้Treeแต่ละต้นมีลักษณะไม่เหมือนกัน จากนั้นจะใช้วิธีการทำ Majority Vote ช่วยในการตัดสินผลการทำนาย โดยนำผลการทำนายของTreeแต่ละต้นมาร่วมกันตัดสินใจ จากนั้นจึงเลือกผลลัพธ์การทำนายที่ได้รับผลโหวตมากที่สุด



ภาพประกอบ 10 แสดงการทำงานของ อัลกอริทึมRandom Forest

ที่มา: Avinash Navlani. (2018). Understanding Random Forest s Classifiers in Python. Retrieved from https://www.datacamp.com/community/tutorials/random-forests-classifier-python

สำหรับ Random Forest ที่ใช้ในงานวิจัยนี้มีการปรับค่าพารามิเตอร์หลักในการสร้างแบบจำลองทำนายแบบ Random Forest ได้แก่

min_samples_split	คือ จำนวนตัวอย่างขั้นต่ำที่ใช้ในการแบ่งโหนดภายใน
max_leaf_nodes	คือ จำนวนโหนดสูงสุดที่ใช้จัดการค่า impurity
criterions	คือ ฟังก์ชันก์สำหรับการวัดคุณภาพของการแบ่งข้อมูล ซึ่งได้แก่ Gini และ Entropy
max_depth	คือ จำนวนลำดับชั้นของต้นไม้แต่ละต้น

การวัดประสิทธิภาพของอัลกอริทึม

ในส่วนของการทำนายแบบการจำแนกประเภท (Classification) ในงานวิจัยนี้จะใช้ค่า Accuracy, Precision, Recall, Macro F1 และ Confusion Matrix ในการวัดประสิทธิภาพของอัลกอริทึม

1. Accuracy คือการวัดค่าความถูกต้องโดยรวมของระบบระหว่างค่าจริงและค่าการทำนาย ถ้าหากค่า Accuracy มีค่ามาก นั่นหมายถึงค่าการทำนายนั้นสามารถทำนายได้ถูกต้องใกล้เคียงกับค่าจริง ดังสมการที่ 2

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (2)$$

จากสมการ

Accuracy	คือ ค่าความถูกต้อง
TP	คือ ค่าการทำนายที่ทำนายว่าจริงซึ่งตรงกับค่าจริง
TN	คือ ค่าการทำนายที่ทำนายว่าไม่จริงซึ่งตรงกับค่าจริง
FP	คือ ค่าการทำนายที่ทำนายว่าไม่จริงซึ่งไม่ตรงกับค่าจริง
FN	คือ ค่าการทำนายที่ทำนายว่าจริงซึ่งไม่ตรงกับค่าจริง

2. Precision คือการวัดค่าความแม่นยำโดยวัดจากความซ้ำเดิมของค่าการทำนายที่ทำนายได้ถูกต้องตรงกับค่าจริง หากค่า Precision มีค่ามาก นั่นหมายถึงค่าการทำนายนั้นสามารถทำนายได้แม่นยำใกล้เคียงกับค่าจริง ดังสมการที่ 3

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

จากสมการ

Precision คือ ค่าความแม่นยำ

TP คือ ค่าการทำนายที่ทำนายว่าจริงซึ่งตรงกับค่าจริง

FP คือ ค่าการทำนายที่ทำนายว่าไม่จริงซึ่งไม่ตรงกับค่าจริง

3. Recall คือการวัดค่าความครบถ้วน ซึ่งหมายถึงอัตราส่วนการวัดค่าการทำนายที่ทำนายได้ถูกต้องตรงกับค่าจริงจากจำนวนของค่าจริงทั้งหมด หากค่า Recall มีค่ามาก นั่นหมายถึงค่าการทำนายนั้นสามารถทำนายได้อย่างครบถ้วนใกล้เคียงกับค่าจริง ดังสมการที่ 4

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

จากสมการ

Precision คือ ค่าความแม่นยำ

TP คือ ค่าการทำนายที่ทำนายว่าจริงซึ่งตรงกับค่าจริง

FN คือ ค่าการทำนายที่ทำนายว่าจริงซึ่งไม่ตรงกับค่าจริง

4. Macro F1 คือการวัดค่าประสิทธิภาพโดยเฉลี่ยของค่า F1 โดยที่ค่า F1 คือการวัดค่าเฉลี่ยการทำนายระหว่างค่า Precision และค่า Recall ดังสมการที่ 5

$$\text{Macro F1} = \left(\frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}\right)/\text{N} \tag{5}$$

จากสมการ

Macro F1 คือ ค่าประสิทธิภาพโดยเฉลี่ยของค่า F1

Precision คือ ค่าความแม่นยำ

Recall คือ ค่าความครบถ้วน

5. Confusion Matrix คือ การเปรียบเทียบประสิทธิภาพของอัลกอริทึมระหว่างค่าการทำนายและค่าจริงในรูปแบบของตาราง ดังนี้

	Actually Positive	Actually Negative
Predicted Positive	TP	FP
Predicted Negative	FN	TN

จากตาราง

Predicted Positive คือ ค่าการทำนายเชิงบวก

Predicted Negative คือ ค่าการทำนายเชิงลบ

Actually Positive คือ ค่าจริงเชิงบวก

Actually Negative คือ ค่าจริงเชิงลบ

TP คือ ค่าการทำนายที่ทำนายว่าจริงซึ่งตรงกับค่าจริง

TN คือ ค่าการทำนายที่ทำนายว่าไม่จริงซึ่งตรงกับค่าจริง

FP คือ ค่าการทำนายที่ทำนายว่าไม่จริงซึ่งไม่ตรงกับค่าจริง

FN คือ ค่าการทำนายที่ทำนายว่าจริงซึ่งไม่ตรงกับค่าจริง

งานวิจัยที่เกี่ยวข้อง

การทบทวนวรรณกรรมของงานวิจัยนี้ได้ทำการศึกษางานวิจัยที่เกี่ยวข้องกับการทำนายผลการเรียนของนักเรียน มีรายละเอียดดังต่อไปนี้

(1) บทความวิจัยเรื่อง Machine Learning Based Student Grade Prediction: A Case Study โดย Zafar Iqbal , Juniad Qadir , Adnan Noor Mian และ Faisal Kamiran (Iqbal, Qadir, Mian, และ Kamiran, 2017)

งานวิจัยนี้นำเสนอการใช้ Machine Learning ประเภท regression ชนิด Collaborative Filtering (CF), Matrix Factorization (MF) และ Restricted Boltzmann Machine (RBM) ในการทำนายเกรดของนักเรียนระดับปริญญาตรีจำนวน 255 คน จากภาควิชาวิศวกรรมไฟฟ้า มหาวิทยาลัย Information Technology University (ITU) ประเทศปากีสถาน ซึ่งใช้ข้อมูลการลงทะเบียนเรียน ข้อมูลคะแนนนักเรียนที่ใช้สำหรับสอบเข้ามหาวิทยาลัย และข้อมูลผลการเรียนของนักเรียน จากผลการทดลองพบว่า โมเดล RBM มีความแม่นยำสูงสุดในการทำนายเกรดของนักเรียนโดยมีค่า Root Mean Square Error (RMSE) เท่ากับ 0.3

(2) บทความวิจัยเรื่อง Feature Extraction For Classifying Students Based On Their Academic Performance โดย Agoritsa Polyzou และ George Karypis (Polyzou และ Karypis, 2018)

งานวิจัยประเภท Classification นี้นำเสนอการใช้ Decision Tree (DT) , Linear Support Vector Machine, Random Forest และ Gradient Boosting ในการแยกคุณลักษณะสำหรับการจำแนกนักเรียนตามประสิทธิภาพทางด้านวิชาการของนักเรียนระดับปริญญาตรีมหาวิทยาลัย Minnesota ประเทศสหรัฐอเมริกา เพื่อทำนายความเสี่ยงที่จะล้มเหลวในการเรียน โดยใช้ feature 8 ชนิด ได้แก่ ข้อมูลผลการเรียนของนักเรียน,ข้อมูลสถานภาพของนักเรียน, ภาระงานของนักเรียน, ความยากและความนิยมของรายวิชา, ความรู้พื้นฐานของรายวิชาและสาขาวิชาที่เรียน, ข้อมูลความสามารถอื่นๆ, ระดับความสามารถของนักเรียนกับรายวิชาที่เรียนและข้อมูลรายวิชาเฉพาะสาขา จากผลการทดลองพบว่า เมธอดที่ดีที่สุดในการทำนายผล คือ Gradient Boosting และ Random Forests โดยวัดจากค่า AUC และ F1

(3) บทความวิจัยเรื่อง Next-Term Student Performance Prediction: A Recommender Systems Approach โดย Mack Sweeney, Huzefa Rangwala, Jaime Lester และ Aditya Johri (Sweeney, Rangwala, Lester, และ Johri, 2016)

งานวิจัยประเภท Regression นี้แบ่งงานสำหรับการวัดผลการพัฒนาระบบการทำนายเกรดของนักศึกษาสำหรับการลงทะเบียนเรียนในภาคเรียนถัดไป ซึ่งแบ่งระดับของเกรดออกเป็น 5 ระดับ(0 – 4) โดยใช้ข้อมูลของนักศึกษาระดับปริญาตรีหลักสูตร 4 ปี การวิจัยแบ่งออกเป็น 3 แนวทาง ได้แก่ แนวทางที่ 1 Simple baselines ใช้วิธี Uniform Random (UR) , Global Mean และ Mean of Means แนวทางที่ 2 Matrix Factorization Methods ใช้วิธี Singular Value Decomposition, SVD-KNN: SVD Post-Processed With KNN และ Factorization Machine แนวทางที่ 3 Common Regression Models ใช้วิธี Random Forest, Stochastic Gradient Descent Regression , K-Nearest Neighbors และ Personalized Multi-Linear Regression

พบว่า แบบจำลอง Factorization Machine และ Random Forest ให้ผลลัพธ์ดีที่สุดในกรณีที่มีจำนวนข้อมูลการเรียนรู้ไม่เพียงพอต่อความจำเป็นของแบบจำลอง(cold-start record) และกรณีที่ไม่มีปัญหาด้านจำนวนข้อมูลไม่เพียงพอต่อความจำเป็นของแบบจำลอง(non cold-start record) ซึ่งใช้ข้อมูลในการทำนายจากการหา feature importance ได้แก่ student bias, course bias, instructor bias และ course discipline โดยมีค่า Root Mean Square Error เท่ากับ 0.7709 และ 0.7775 ตามลำดับ จากผลลัพธ์ดังกล่าวพบว่าผู้วิจัยปรับปรุงการทำนายโดย

ใช้แบบจำลอง hybrid ระหว่างแบบจำลอง Factorization Machine และ Random Forest โดยให้ค่า Root Mean Square Error เท่ากับ 0.7443

(4) บทความวิจัยเรื่อง Predicting Student Grades using Machine Learning โดย Naveen Venkat , Sahaj Srivastava และ Lakshya Garg (Venkat, Srivastava, และ Garg, 2018)

งานวิจัยนี้นำเสนอการทำนายเกรดของนักเรียนประเภท Classification โดยใช้เทคนิค Decision Tree, Naive Bayes , SVM และ K-Nearest Neighbor ซึ่งงานวิจัยนี้ทดลองทำการทำนาย 3 ประเภท ดังนี้ คือ (1)ทำนายผลการเรียนโดยใช้กลุ่มคะแนนสอบเท่านั้น ได้แก่ Midterm semester, Quiz 1, Quiz 2, Part A score, Part B score (2)ทำนายผลการเรียนโดยใช้ข้อมูลเพิ่มเติม ได้แก่ Year, Attendance, CGPA และ Midterm semester Collection (3)ทำนายผลการเรียนโดยใช้ข้อมูลทั้งหมด(ข้อมูลคะแนนสอบและข้อมูลเพิ่มเติม) ผลการทดลองพบว่า Linear-SVM ให้ผลการทำนายดีที่สุดจากค่า mean ที่ 0.88, ค่า standard deviation ที่ 0.08 ในการทดลองประเภทที่ 1 และในการทดลองประเภทที่ 2 ได้ค่า mean ที่ 0.56, ค่า standard deviation ที่ 0.21 สุดท้ายในการทดลองประเภทที่ 3 ได้ค่า mean ที่ 0.75, ค่า standard deviation ที่ 0.14 ตามลำดับ

(5) บทความวิจัยเรื่อง Predicting Grades โดย Yannick Meier , Jie Xu , Onur Atan และ Mihaela van der Schaar (Meier, Xu, Atan, และ Schaar, 2016)

งานวิจัยนี้นำเสนออัลกอริทึมที่พัฒนาขึ้นสำหรับทำนายผลการเรียนของนักเรียนในหลักสูตรการประมวลผลสัญญาณดิจิตอลระดับปริญญาตรี มหาวิทยาลัยแคลิฟอร์เนียลอสแอนเจลิส จำนวน 700 คน ข้อมูลที่ใช้เป็นข้อมูลในช่วง 7 ปีการศึกษา ประกอบด้วย การบ้าน 20% การสอบกลางภาค 25% การสอบปลายภาค 40% และโปรเจ็คงาน 15% งานวิจัยเปรียบเทียบการทำนายผลการเรียนโดยใช้อัลกอริทึมที่พัฒนาโดยใช้เทคนิค Neighborhood Radius ร่วมกับเทคนิค Brute force เปรียบเทียบกับ Benchmarks โดยใช้ nearest neighbor method สำหรับงาน Classification ซึ่งใช้ accuracy , precision , recall และ false positive/negative rate ในการวัดประสิทธิภาพ และใช้ linear/logistic regression สำหรับงาน Regression ซึ่งใช้ Average Absolute Prediction Errors ร่วมกับ Average Prediction Time ในการวัดประสิทธิภาพการทำนายผลการเรียนใน 5 วิธีการ ดังนี้ (1) การทำนายผลการเรียนจากคะแนนสอบล่าสุดเปรียบเทียบกับ Benchmarks ในงาน regression พบว่าอัลกอริทึมที่นำเสนอมีข้อผิดพลาดในการทำนายน้อยกว่า linear regression ถึง 65% (2) การทำนายผลการเรียนจาก

คะแนนระหว่างปีการศึกษาที่มีผู้สอนที่แตกต่างกันในแต่ละปีในงาน regression พบว่า การทำนายโดยใช้ผลการเรียนที่ได้จากการสอนโดยผู้สอนคนเดียวกันมีความแม่นยำมากกว่าข้อมูลผลการเรียนที่ได้จากการสอนโดยผู้สอนแตกต่างกันเพียงเล็กน้อย (3) การทำนายผลการเรียนจากหลักสูตรก่อนหน้า ในงาน regression พบว่า ข้อมูลคะแนนสอบในชั้นเรียนนั้นให้ผลการทำนายดีกว่าการทำนายด้วยข้อมูลผลการเรียนจากการบ้าน (4) การทำนายผลการเรียนเปรียบเทียบกับ Benchmarks ในงาน classification พบว่า อัลกอริทึมที่นำเสนอให้ค่า accuracy , precision และ recall ดีกว่า logistic regression (5) การทำนายผลการเรียนในงาน classification ด้วยอัลกอริทึมที่นำเสนอให้ค่าความแม่นยำที่ 76% และสะสมต่อเนื่องที่ 80%

(6) บทความวิจัยเรื่อง Predicting Students' GPA and Developing Intervention Strategies Based on Self-Regulatory Learning Behaviors โดย Amin Zollanvari , Refik Caglar Kizilirmark , Yau Hee Kho และ Daniel Hernández-Torrano (Zollanvari, Kizilirmak, Kho, และ Hernández-Torrano, 2017)

งานวิจัยประเภท Classification นำเสนอการสร้างชุดแบบทดสอบพฤติกรรมการเรียนรู้ด้วยตนเองที่สามารถใช้เป็น feature ในการทำนายเกรดของนักเรียนได้ โดยใช้ข้อมูลจากการทำแบบทดสอบของนักศึกษาชั้นปีที่ 2-4 ที่ลงทะเบียนเรียนเฉพาะในโปรแกรมวิศวกรรมไฟฟ้าที่ Nazarbayev University ได้แก่ การจัดการเวลาในการเรียนรู้, สภาพแวดล้อมการศึกษา, ทักษะด้านการเตรียมสอบ, ทักษะการจดบันทึก,การอ่านและการเขียน, เวลาเรียน และเกรดเฉลี่ยสะสม โดยใช้เทคนิค Maximum - Weight First - Order Dependence Tree ในการทำนายผลการเรียนและใช้ Confusion Matrix และค่า Accuracyในการประเมินความถูกต้องของโมเดล พบว่า โมเดลมีผล Accuracy ที่ 82%

(7) บทความวิจัยเรื่อง การใช้เทคนิคการทำเหมืองข้อมูลในการจำแนกและคัดเลือกแขนงวิชาสำหรับนักศึกษาคณะเทคโนโลยีสารสนเทศ โดย จิราภา เลาหะวรนันท์, รชต ลิ้มสุทธิวันภูมิ, บัณฑิต ฐานะโสภณ และ พรฤดี เนติโสภากุล (เลาหะวรนันท์, ลิ้มสุทธิวันภูมิ, ฐานะโสภณ, และ เนติโสภากุล, 2018)

งานวิจัยประเภท Classification นี้นำเสนอการพัฒนาระบบแนะนำแขนงวิชาสำหรับนักศึกษาคณะเทคโนโลยีสารสนเทศในปีการศึกษา 2555-2557 สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ซึ่งแขนงวิชาที่แนะนำได้แก่ แขนงวิชาวิศวกรรมซอฟต์แวร์ (Software) , แขนงวิชาเทคโนโลยีเครือข่ายและระบบ(Network and System Technology), แขนงวิชาการพัฒนาสื่อประสมและเกมส์ (Multimedia) และแขนงวิชาอัจฉริยะทางธุรกิจ(Business

Intelligence) ซึ่งใช้ข้อมูลเกรดรายวิชาพื้นฐานจำนวน 14 รายวิชา, คะแนนแบบทดสอบความถนัดเฉพาะด้าน และเพศ โดยใช้เทคนิค Ensemble Vote เพื่อสร้างโมเดลจากเทคนิคการทำนาย 5 เทคนิค ได้แก่ เทคนิค Decision Tree , Naïve Bayesian , Neural Network และ Logistic Regression โดยใช้ Confusion Matrix รวมถึง Precision และ Recall ในการวัดความแม่นยำของโมเดล พบว่าการทำนายผลมีความแม่นยำอยู่ที่ 72.92% โดยแขนงวิชา Software ได้ค่าความแม่นยำสูงสุดที่ 86.67% และแขนงวิชาที่ได้ค่าความแม่นยำน้อยที่สุดคือ แขนงวิชา Business Intelligence ที่ 60.83% ซึ่งค่าความแม่นยำของแขนงวิชาที่ต่างกันนั้นน่าจะมีผลมาจากลักษณะของผู้เรียน ซึ่งผู้เรียนแขนงวิชา Software มีลักษณะคล้ายกันคือได้คะแนนเต็มด้านการเขียนโปรแกรม ซึ่งทำให้การทำนายง่ายกว่าแขนงวิชา Business Intelligence ที่ผู้เรียนแต่ละคนมีลักษณะที่ต่างกัน ข้อจำกัดของโมเดลคือต้องการข้อมูลปริมาณที่มากขึ้นหรือปัจจัยที่บ่งเฉพาะความถนัดของผู้เรียนแต่ละแขนงวิชาเพื่อลดความแตกต่างของข้อมูลและเพิ่มความแม่นยำของโมเดล

(8) บทความวิจัยเรื่อง ระบบทำนายผลการเรียนนักศึกษาออนไลน์โดยใช้เคเนียเร็ซเนเบอะ (Online Student Forecast System By Using K-Nearest Neighbor) โดย กริช สมกันธา, วิไลพร กุลตังวัฒนา, ธีระวัฒน์ หัสโก และ จิระพงศ์ รอดชมภู (สมกันธา, กุลตังวัฒนา, หัสโก, และ รอดชมภู, 2532)

งานวิจัยประเภท Classification นี้นำเสนอระบบทำนายผลการเรียนนักศึกษาออนไลน์เพื่อลดปัญหาความผิดพลาดในด้านการให้คำปรึกษาของอาจารย์ที่ปรึกษาเพื่อตัดสินใจถอดถอนรายวิชาเรียนของนักศึกษา ซึ่งใช้เทคนิค K-Nearest Neighbor โดยเปรียบเทียบประสิทธิภาพกับวิธีการของเบย์และจากผู้เชี่ยวชาญ และวัดค่าความถูกต้องด้วย Confusion Matrix สำหรับข้อมูลคุณลักษณะใช้เทคนิค Feature Extraction ในการสกัดคุณลักษณะเด่น 6 คุณลัษณะ ได้แก่ คะแนนสอบย่อย, ความถี่ในการเข้าชั้นเรียน, คะแนนสอบกลางภาค, คะแนนเก็บภาคปฏิบัติ, คะแนนการบ้าน, คะแนนภาระงาน และคะแนนรวมตัดเกรด จากข้อมูลนักศึกษาวิชาพื้นฐานระบบเทคโนโลยีสารสนเทศและวิชาระบบคอมพิวเตอร์และสถาปัตยกรรมจำนวน 50 คน พบว่าเทคนิค K-Nearest Neighbor ให้ผลการทำนายที่ดีที่สุดเมื่อ K มีค่าเท่ากับ 15-NN ที่ค่าความแม่นยำ 90% เปรียบเทียบกับกับเทคนิคของเบย์ที่ความค่าความแม่นยำอยู่ที่ 84% และความแม่นยำจากผู้เชี่ยวชาญที่ 82% ซึ่งสรุปได้ว่าโมเดลมีการทำงานได้ดีจากข้อมูลที่มีจำนวนไม่มาก หากมีการเพิ่มปริมาณข้อมูลอาจได้ค่าความแม่นยำที่ต่างออกไป

(9) บทความวิจัยเรื่อง การพยากรณ์โอกาสสำเร็จการศึกษาของนักศึกษาโดยใช้ซัพพอร์ตเวกเตอร์แมชชีน (Graduation Forecasting Using Support Vector Machine) โดย พรรณิภา บุตรเอก และ สุรเดช บุญลือ (บุตรเอก, 2014)

งานวิจัยประเภท Classification นี้นำเสนอการทำนายโอกาสการสำเร็จการศึกษาของนักศึกษา โดยใช้ชุดข้อมูลระเบียนประวัติของนักศึกษาระดับปริญญาตรีชั้นปีที่ 1 หลักสูตร 4 ปี สาขาวิชาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยนอร์ทกรุงเทพ จำนวน 3 ภาคเรียน ระหว่างปีการศึกษา 2547 – 2551 ข้อมูลที่ใช้ในงานวิจัยแบ่งเป็น 2 กลุ่ม ได้แก่ (1)ชุดข้อมูลประวัติส่วนตัว เช่น เพศ, รายได้บิดา, รายได้มารดา, สถานะบิดา, สถานะมารดา, เกรดเฉลี่ยสถานศึกษาเดิม, หมู่เลือด, กลุ่มโรงเรียนที่จบการศึกษา(รัฐ/เอกชน), ภูมิภาคที่อยู่อาศัย (2)ชุดข้อมูลผลการเรียนของนักศึกษา เช่น สถานะการได้รับทุน, กลุ่มการเรียน, จำนวนปีที่กำหนด, เกรดเฉลี่ยภาคเรียนที่ 1, เกรดเฉลี่ยภาคเรียนที่ 2, เกรดเฉลี่ยภาคฤดูร้อน, เกรดเฉลี่ยสะสมภาคเรียนที่ 1, เกรดเฉลี่ยสะสมภาคเรียนที่ 2 และเกรดเฉลี่ยสะสมภาคฤดูร้อน การทำนายผลลัพธ์ใช้เทคนิค Support Vector Machine เทคนิค Decision Tree และเทคนิคโครงข่ายประสาทเทียมแบบย้อนกลับ(BP-ANN) สำหรับเทคนิค Support Vector Machine ใช้เคอร์เนลฟังก์ชันแบบ Polynomial Kernel เนื่องจากให้ค่าความแม่นยำสูงที่สุดเมื่อเปรียบเทียบกับ Normalise Polynomial Kernel , Radial Basis Function Kernel และ Pearson VII function-based universal kernel โดยใช้ตัววัดประสิทธิภาพความแม่นยำ ดังนี้ ค่า Precision , ค่า Recall , ค่า F-Measure และค่า Accuracy พบว่าเทคนิค Support Vector Machine มีผลการทำนายสำหรับค่า Accuracy ที่ 89.13% , ค่า Precision ที่ 0.88 , ค่า Recall 0.89 , ค่า RMSE ที่ 0.33 และค่าประสิทธิภาพโดยรวมที่ 0.86 โดยค่าผลลัพธ์ที่ได้มาจากการทำนายด้วยชุดข้อมูลของนักศึกษาจำนวน 138 คน 18 คุณลักษณะ ซึ่งเป็นจำนวนข้อมูลที่ค่อนข้างน้อย

สรุปงานวิจัยที่ศึกษาทั้งหมด

ลำดับที่	ชื่อเรื่อง	ปีที่พิมพ์	Model					Other
			LR	SVM	KNN	XG boost	RF	
1	Machine Learning Based Student Grade Prediction: A Case Study	2017						NB, CF, MF, RBM
2	Feature Extraction For Classifying Students Based On Their Academic Performance	2018		√			√	GB, DT
3	Next-Term Student Performance Prediction: A Recommender Systems Approach	2016			√		√	SGD, PMLR
4	Predicting Student Grades using Machine Learning	2018		√	√			DT, NB
5	Predicting Grades	2016	√		√			Bays
6	Predicting Students' GPA and Developing Intervention Strategies Based on Self-Regulatory Learning Behaviors	2017						MWDT
7	การใช้เทคนิคการทำเหมืองข้อมูลในการจำแนกและคัดเลือกแขนงวิชาสำหรับนักศึกษาคณะเทคโนโลยีสารสนเทศ	2015	√					DT, NN
8	Online Student Forecast System By Using K-Nearest Neighbor	1989			√			
9	Graduation Forecasting Using Support Vector Machine	2014		√				DT, BP-ANN

ภาพประกอบ 11 แสดงตารางสรุปงานวิจัยที่ศึกษา

บทที่ 3
วิธีการดำเนินการวิจัย

ในการวิจัยครั้งนี้ ผู้วิจัยได้ดำเนินการตามขั้นตอนดังนี้

1. การกำหนดประชากรและกลุ่มตัวอย่าง
2. การสร้างเครื่องมือที่ใช้ในการวิจัย
3. การรวบรวมข้อมูล
4. การจัดกระทำและการวิเคราะห์ผลข้อมูล

การกำหนดประชากรและกลุ่มตัวอย่าง

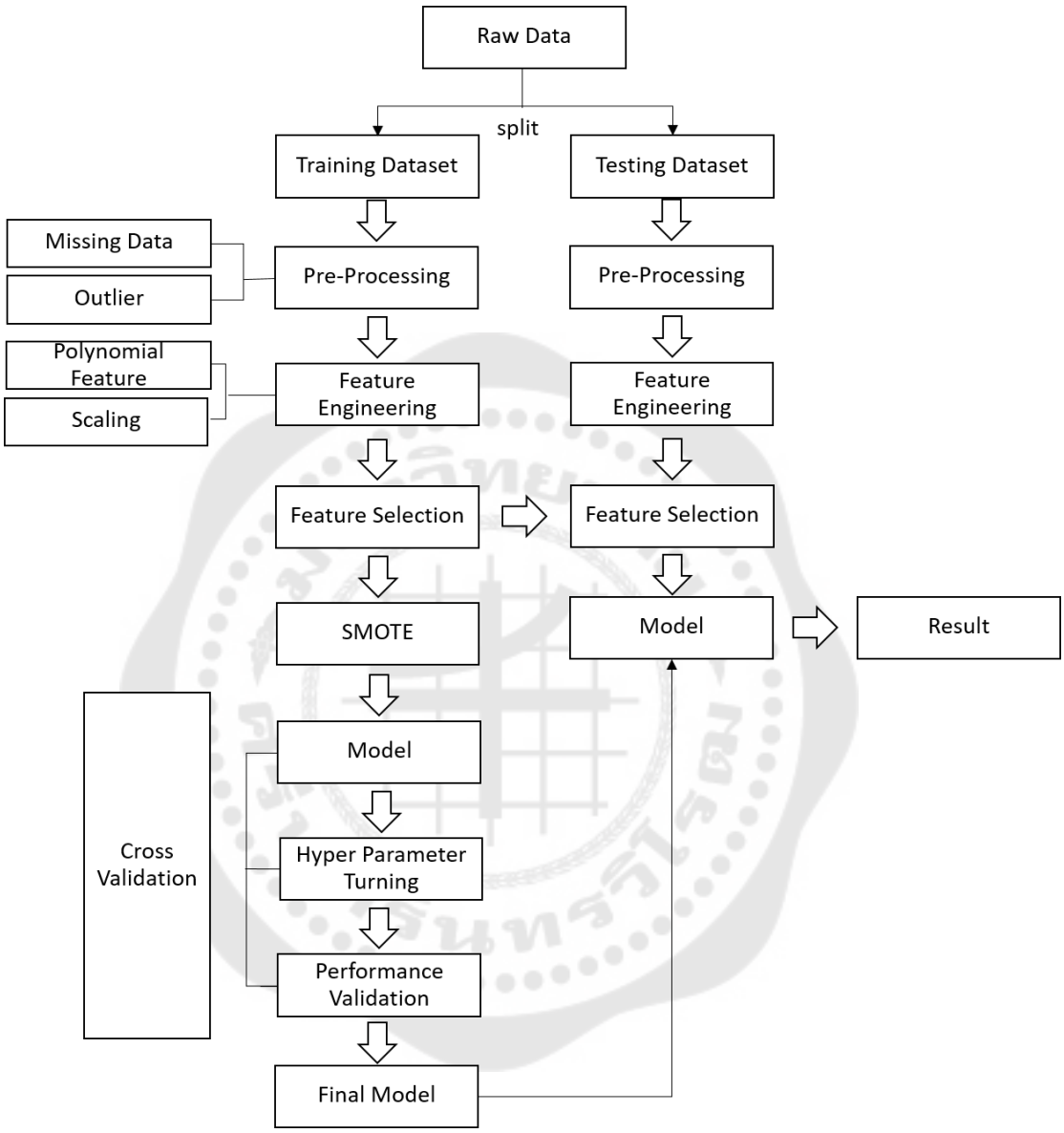
ประชากร

งานวิจัยนี้ได้รับความอนุเคราะห์ข้อมูลผลการเรียนของนักเรียนจากโรงเรียนมัธยมศึกษาแห่งหนึ่ง โดยเก็บข้อมูลผลการเรียนรายบุคคลจำนวน 3 ปีการศึกษา ตั้งแต่ปีการศึกษา 2558 – 2560 ประกอบด้วย ข้อมูลนักเรียนจำนวน 1382 คน และข้อมูล feature ทั้งหมด 15 คอลัมน์

การเลือกกลุ่มตัวอย่าง

การเลือกกลุ่มตัวอย่างจะดำเนินการแบ่งข้อมูลออกเป็น 2 ชุด ได้แก่ ชุดข้อมูลการเรียนรู้ (Training Dataset) และชุดข้อมูลการทดสอบ (Testing Dataset) โดยการเลือกกลุ่มตัวอย่างจะแบ่งข้อมูลในอัตราส่วน 80 : 20

การสร้างเครื่องมือที่ใช้ในการวิจัย



ภาพประกอบ 12 แสดงขั้นตอนการสร้างแบบจำลองเพื่อทำนายผลการเรียนของนักเรียน

ขั้นตอนการสร้างแบบจำลองการทำนายผลการเรียนของนักเรียน โดยเริ่มจากการแบ่ง training dataset และ testing dataset ในส่วนของ training dataset จะทำการ clean data ในขั้นตอน pre-processing จากนั้นใช้เทคนิค polynomial feature และ scaling data ในขั้นตอนการทำ feature engineering และทำการเลือก feature ที่สำคัญต่อการ train data โดยอาศัยการทำ feature selection ร่วมกับการหาค่า correlation เมื่อได้ feature จากขั้นตอนดังกล่าวแล้วจะ

ทำการเพิ่มข้อมูลในแต่ละ class โดยใช้เทคนิค SMOTE จากนั้นนำข้อมูลเข้าสู่แบบจำลองเพื่อทำการหาค่าพารามิเตอร์ที่ดีที่สุดในขั้นตอน Hyper Parameter Tuning และวัดประสิทธิภาพการทำงานของแบบจำลองโดยทำ Cross Validation จากนั้นนำแบบจำลองที่ได้ไปใช้ทำนายผลการเรียนในขั้นตอนการ testing โดยใช้มูล testing dataset จากการแบ่งขั้นตอนแรก จากนั้นทำการ clean data ในขั้นตอน pre-processing และใช้เทคนิค polynomial feature และ scaling data ในขั้นตอนการทำ feature engineering แล้วทำการเลือก feature ในขั้นตอนการทำ feature selection ร่วมกับการหาค่า correlation จากนั้นทำการทำนายผลการเรียนของนักเรียนและแสดงผลลัพธ์ที่ได้

การรวบรวมข้อมูล

ข้อมูลผลการเรียนที่ใช้สำหรับงานวิจัยนี้ได้รับความอนุเคราะห์จากการเก็บรวบรวมโดยโรงเรียนระดับมัธยมศึกษาแห่งหนึ่งในจังหวัดสุพรรณบุรี โดยใช้ข้อมูลของนักเรียนชั้นมัธยมศึกษาตอนต้น(ชั้นมัธยมศึกษาปีที่ 1 -3) ตั้งแต่ปีการศึกษา 2558 ถึงปีการศึกษา 2560 ผลการเรียนของนักเรียนจะแบ่งเป็น 3 ระดับ ดังแสดงในตารางที่ 1

ตาราง 1 แสดงรายละเอียดเกณฑ์ระดับคุณภาพผลการเรียน

ช่วงคะแนน	ระดับคุณภาพผลการเรียน
60-69	พอใช้
70-79	ดี
80 ขึ้นไป	ดีมาก

ซึ่งข้อมูลมีจำนวนทั้งหมด 1382 รายการ 15 คอลัมน์ ดังแสดงในตารางที่ 2

ตาราง 2 แสดงรายละเอียดของคุณลักษณะ(Feature) ซึ่งเป็นข้อมูลตัวเลขที่มีค่าอยู่ในช่วง 0-100 โดยที่ i = 1…5

feature	คำอธิบาย
M[i]	คะแนนรายวิชาคณิตศาสตร์ที่ i
SC[i]	คะแนนรายวิชาวิทยาศาสตร์ที่ i
EN[i]	คะแนนรายวิชาภาษาอังกฤษที่ i

การจัดกระทำและการวิเคราะห์ผลข้อมูล

1. ทำการแบ่งข้อมูลเป็น train dataset และ test dataset ในแต่ละรายวิชา จากนั้นทำการจัดรูปแบบ(Format)ให้เหมาะสมและทำการเพิ่ม feature โดยการหาค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของแต่ละรายวิชา

2. ศึกษาและวิเคราะห์ข้อมูลใน dataset โดยจัดการกับข้อมูล missing value, ข้อมูลoutlier และหาค่าความสัมพันธ์ระหว่างข้อมูล(Correlation) จากนั้นทำFeature Engineering ได้แก่ การใช้ Polynomial Feature และ Scaling เพื่อสร้าง feature ใหม่ขึ้นมาและจัดการค่าข้อมูลให้มีค่าอยู่ในช่วงเดียวกัน

3. ใช้เทคนิค RFE เพื่อเลือก feature ที่มีความสำคัญกับ Label จากนั้นใช้เทคนิค SMOTE จัดการกับ Imbalance dataset

4. สร้างแบบจำลองในการทำนายผลการเรียน และทำการหาค่าพารามิเตอร์ที่ดีที่สุดในการทำนายโดยใช้เทคนิค Hyperparameter Selection หรือ Grid Search

4.1 สร้างแบบจำลองโดยใช้อัลกอริทึม XGBoost

4.2 สร้างแบบจำลองโดยใช้อัลกอริทึม logistic regression

4.3 สร้างแบบจำลองโดยใช้อัลกอริทึม SVM

4.4 สร้างแบบจำลองโดยใช้อัลกอริทึม KNN

4.5 สร้างแบบจำลองโดยใช้อัลกอริทึม Random Forest

5. เครื่องมือที่ใช้วัดประสิทธิภาพของแบบจำลอง

5.1 Confusion matrix

5.2 Accuracy

5.3 Precision

5.4 Recall

5.5 Macro F1-Score

6. สำหรับการดำเนินการทดลองจะกระทำทั้งหมด 4 ครั้ง ดังนี้

6.1 การทดลองครั้งที่ 1 ใช้ feature คะแนนรายวิชาคณิตศาสตร์ 1, คณิตศาสตร์ 2, วิทยาศาสตร์ 1, วิทยาศาสตร์ 2, ภาษาอังกฤษ 1 และภาษาอังกฤษ 2 ในส่วนของ Label คือ ผลการเรียนรายวิชาคณิตศาสตร์ 3, วิทยาศาสตร์ 3 และภาษาอังกฤษ 3

6.2 การทดลองครั้งที่ 2 ใช้ feature คะแนนรายวิชาคณิตศาสตร์ 1, คณิตศาสตร์ 2, คณิตศาสตร์ 3, วิทยาศาสตร์ 1, วิทยาศาสตร์ 2, วิทยาศาสตร์ 3, ภาษาอังกฤษ 1, ภาษาอังกฤษ 2 และภาษาอังกฤษ 3 ในส่วนของ Label คือ ผลการเรียนรายวิชาคณิตศาสตร์ 4, วิทยาศาสตร์ 4 และภาษาอังกฤษ 4

6.3 การทดลองครั้งที่ 3 ใช้ feature คะแนนรายวิชาคณิตศาสตร์ 1, คณิตศาสตร์ 2, คณิตศาสตร์ 3, คณิตศาสตร์ 4, วิทยาศาสตร์ 1, วิทยาศาสตร์ 2, วิทยาศาสตร์ 3, วิทยาศาสตร์ 4, ภาษาอังกฤษ 1, ภาษาอังกฤษ 2, ภาษาอังกฤษ 3 และภาษาอังกฤษ 4 ในส่วนของ Label คือ ผลการเรียนรายวิชาคณิตศาสตร์ 5, วิทยาศาสตร์ 5 และภาษาอังกฤษ 5

6.4 การทดลองครั้งที่ 4 ใช้ feature คะแนนรายวิชาคณิตศาสตร์ 1, คณิตศาสตร์ 2, คณิตศาสตร์ 3, คณิตศาสตร์ 4, คณิตศาสตร์ 5, วิทยาศาสตร์ 1, วิทยาศาสตร์ 2, วิทยาศาสตร์ 3, วิทยาศาสตร์ 4, วิทยาศาสตร์ 5, ภาษาอังกฤษ 1, ภาษาอังกฤษ 2, ภาษาอังกฤษ 3, ภาษาอังกฤษ 4 และภาษาอังกฤษ 5 ในส่วนของ Label คือ ผลการเรียนรายวิชาคณิตศาสตร์ 6, วิทยาศาสตร์ 6และภาษาอังกฤษ 6

สำหรับการทดลองการทำนายนั้นได้ทำการเพิ่มคุณลักษณะโดยการหาค่าเฉลี่ยของคะแนนรายกลุ่มวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษ เพื่อเพิ่มจำนวนคุณลักษณะสำหรับทำนายซึ่งอาจมีความสำคัญกับ Label ผลการเรียนในแต่ละรายวิชา ดังแสดงในตารางที่ 3

ตาราง 3 แสดงรายละเอียดการเพิ่มคุณลักษณะโดยการหาค่าเฉลี่ย ซึ่งเป็นข้อมูลตัวเลขที่มีค่าอยู่ในช่วง 0-100 โดยที่ i = 1...5, และ j>i

feature	คำอธิบาย
ave_M[i-j]	ค่าเฉลี่ยสะสมคะแนนรายวิชาคณิตศาสตร์ M[i], M[i+1],...M[j]
ave_SC[i-j]	ค่าเฉลี่ยสะสมคะแนนรายวิชาวิทยาศาสตร์ SC[i], SC[i+1],...SC[j]
ave_EN[i-j]	ค่าเฉลี่ยสะสมคะแนนรายวิชาภาษาอังกฤษ EN[i], EN[i+1],...EN[j]

	math1	math2	math3	math4	math5	ave5_math	ave4_math	ave3_math	ave2_math	stdv5_math	stdv4_math	stdv3_math	stdv2_math
0	76	75	57	65	68	68.2	66.25	63.33	66.5	7.79	7.46	5.69	2.12
1	76	83	71	78	70	75.6	75.50	73.00	74.0	5.32	6.14	4.36	5.66
2	77	80	77	73	65	74.4	73.75	71.67	69.0	5.81	6.50	6.11	5.66
3	70	77	60	68	63	67.6	67.00	63.67	65.5	6.58	7.44	4.04	3.54
4	70	73	57	68	63	66.2	65.25	62.67	65.5	6.30	6.85	5.51	3.54

ภาพประกอบ 13 แสดงตัวอย่างข้อมูลของ Dataset ผลการเรียนของนักเรียนระดับชั้นมัธยมศึกษาตอนต้นในโปรแกรม

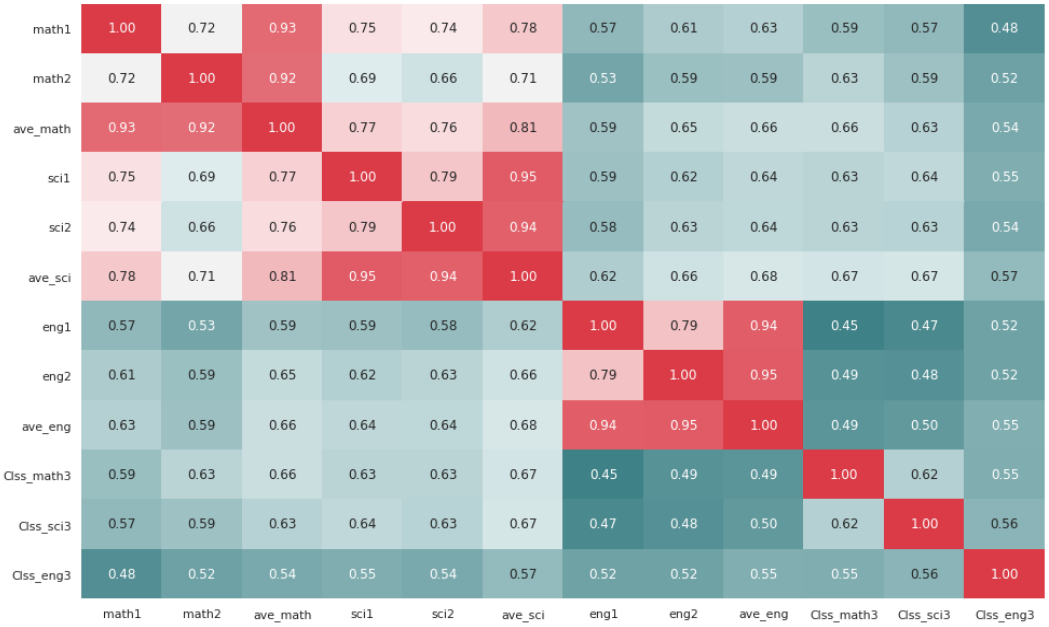
จากรูปภาพที่ 13 แสดงตัวอย่างข้อมูลผลการเรียนของนักเรียนใน Dataset จากโรงเรียนระดับมัธยมศึกษาแห่งหนึ่ง ซึ่งมีข้อมูลจำนวน 1382 รายการ 39 คอลัมน์ โดยข้อมูลประกอบด้วยประเภท integer ได้แก่ math1, math2, math3, math4, math5, sci1, sci2, sci3, sci4, sci5, eng1, eng2, eng3, eng4, eng5, Clss_math3, Clss_sci3, Clss_eng3, Clss_math4, Clss_sci4 Clss_eng4, Clss_math5, Clss_sci5, Clss_eng5, Clss_math6, Clss_sci6 และ Clss_eng6 ข้อมูลประเภท float ได้แก่ ave5_math, ave4_math, ave3_math, ave2_math, ave5_sci, ave4_sci, ave3_sci, ave2_sci, ave5_eng, ave4_eng, ave3_eng และ ave2_eng

ทำการแบ่งข้อมูลเป็น train dataset และ test dataset ของแต่ละรายวิชาดังแสดงในตารางที่ 4

ตาราง 4 แสดงรายละเอียดการแบ่งจำนวนข้อมูล train dataset และ test dataset

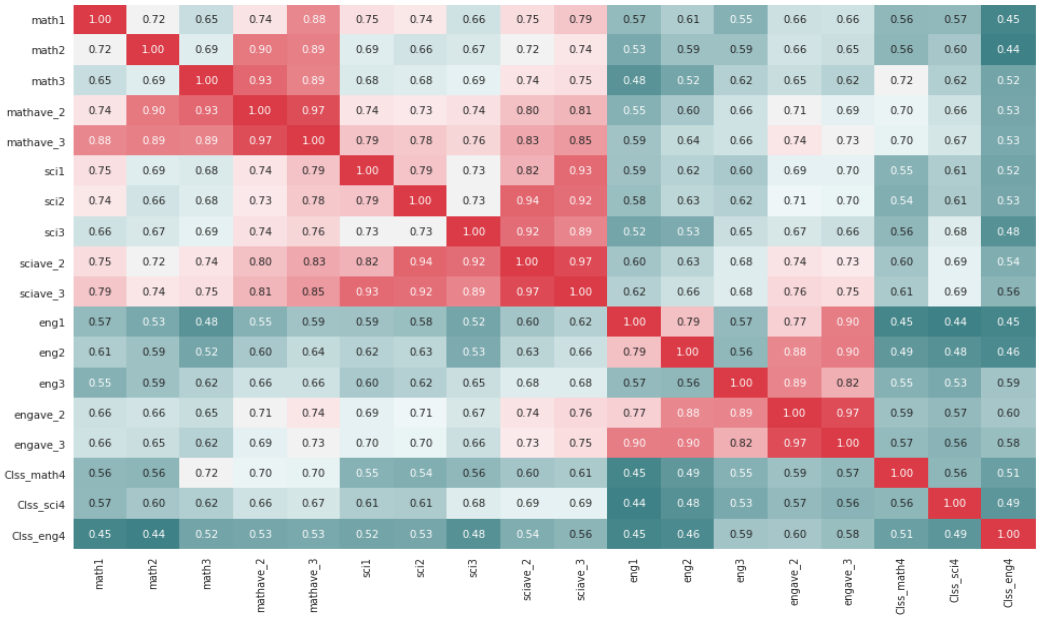
การทดลอง	รายวิชา	train dataset	test dataset	Label	feature
ครั้งที่ 1	คณิตศาสตร์	1068	267	Clss_math3	math1, math2, sci1, sci2, eng1, eng2, ave2_math, ave2_sci, ave2_eng
	วิทยาศาสตร์	1083	271	Clss_sci3	
	ภาษาอังกฤษ	1038	260	Clss_eng3	
ครั้งที่ 2	คณิตศาสตร์	1068	267	Clss_math4	math1, math2, math3, sci1, sci2, sci3, eng1, eng2, eng3, ave3_math, ave2_math, ave3_sci, ave2_sci, ave3_eng, ave2_eng
	วิทยาศาสตร์	1083	271	Clss_sci4	
	ภาษาอังกฤษ	1038	260	Clss_eng4	
ครั้งที่ 3	คณิตศาสตร์	1068	267	Clss_math5	math1, math2, math3, math4, sci1, sci2, sci3, sci4, eng1, eng2, eng3, eng4, ave4_math, ave3_math, ave2_math, ave4_sci, ave3_sci, ave2_sci, ave4_eng, ave3_eng, ave2_eng
	วิทยาศาสตร์	1083	271	Clss_sci5	
	ภาษาอังกฤษ	1038	260	Clss_eng5	
ครั้งที่ 4	คณิตศาสตร์	1068	267	Clss_math6	math1, math2, math3, math4, math5, sci1, sci2, sci3, sci4, sci5, eng1, eng2, eng3, eng4, eng5, ave5_math, ave4_math, ave3_math, ave2_math, ave5_sci, ave4_sci, ave3_sci, ave2_sci, ave5_eng, ave4_eng, ave3_eng, ave2_eng
	วิทยาศาสตร์	1083	271	Clss_sci6	
	ภาษาอังกฤษ	1038	260	Clss_eng6	

ทำการศึกษาและจัดการกับข้อมูลก่อนสร้างแบบจำลอง(Pre-Process) โดยการวิเคราะห์ข้อมูล ได้แก่ การจัดการกับ missing value , การจัดการ outlier , การหาค่าความสัมพันธ์ระหว่างข้อมูล(Correlation)



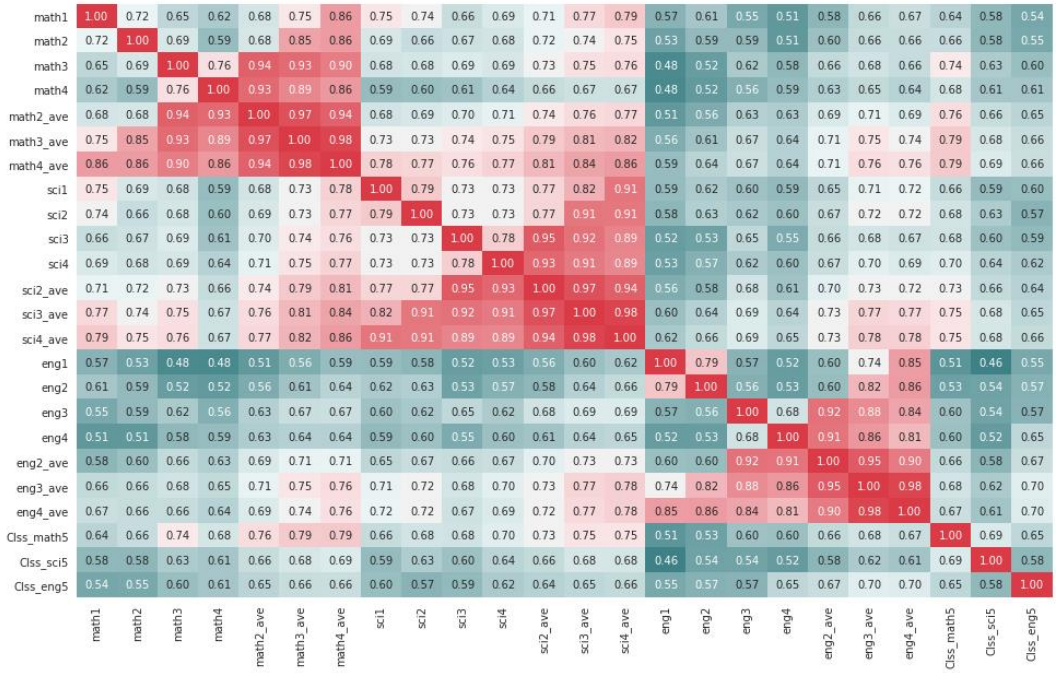
ภาพประกอบ 14 แสดงตัวอย่างการหาค่าความสัมพันธ์(Correlation)
ระหว่าง Feature สำหรับการทดลองครั้งที่ 1

จากรูปภาพที่ 14 แสดงการหาค่าความสัมพันธ์ระหว่างคุณลักษณะทั้งหมดใน dataset กับ Label ของรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษในรูปแบบ correlation matrix ของการทดลองครั้งที่ 1 พบว่า feature ที่สำคัญ 3 อันดับแรกของการทำนายผลการเรียนวิชาคณิตศาสตร์3 ได้แก่ ผลการเรียนเฉลี่ยวิชาวิทยาศาสตร์1-2ที่ค่า 0.67 ผลการเรียนเฉลี่ยวิชาคณิตศาสตร์1-2 ที่ค่า 0.66 วิชาคณิตศาสตร์2 วิทยาศาสตร์1 และวิทยาศาสตร์2 ที่ค่า 0.63 สำหรับการทำนายผลการเรียนวิชาวิทยาศาสตร์3 ได้แก่ ผลการเรียนเฉลี่ยวิชาวิทยาศาสตร์1-2 ที่ค่า 0.67 วิทยาศาสตร์1 ที่ค่า 0.64 วิทยาศาสตร์2และผลการเรียนเฉลี่ยวิชาคณิตศาสตร์1-2 ที่ค่า 0.63 ตามลำดับ สำหรับการทำนายผลการเรียนวิชาภาษาอังกฤษ3 ได้แก่ ผลการเรียนเฉลี่ยวิชาวิทยาศาสตร์1-2 ที่ค่า 0.57 ผลการเรียนเฉลี่ยวิชาภาษาอังกฤษ1-2 ที่ค่า 0.55 ผลการเรียนเฉลี่ยวิชาคณิตศาสตร์1-2 และวิทยาศาสตร์2 ที่ค่า 0.54



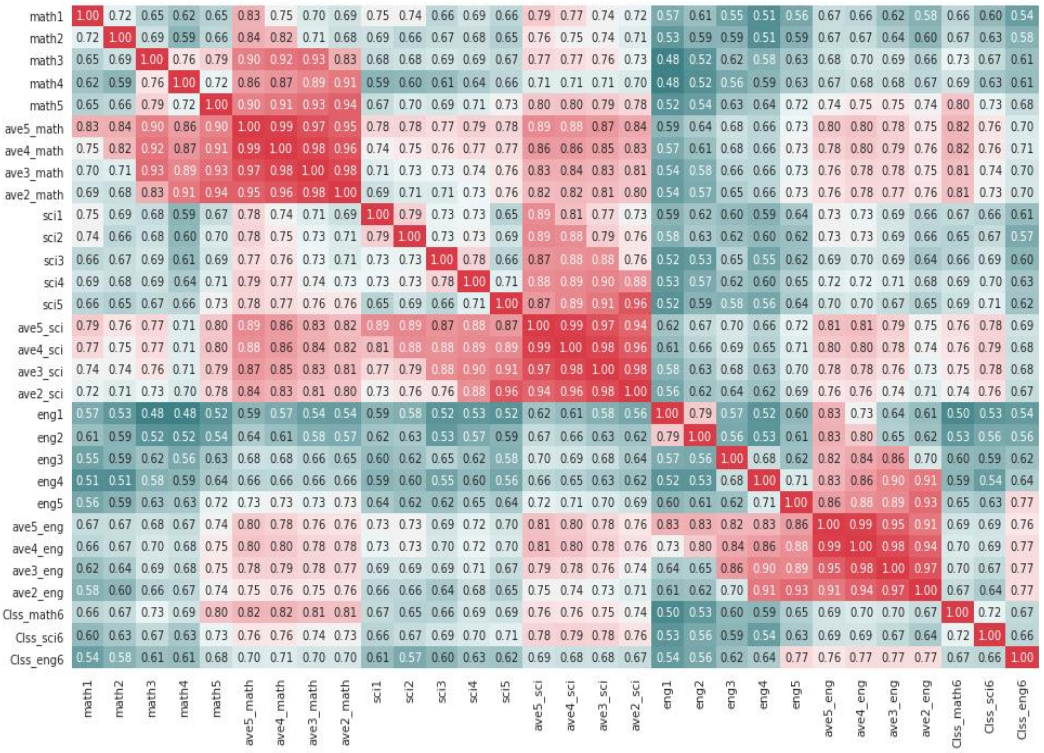
ภาพประกอบ 15 แสดงตัวอย่างการหาค่าความสัมพันธ์(Correlation)
ระหว่าง Feature สำหรับการทดลองครั้งที่ 2

จากรูปภาพที่ 15 แสดงการหาค่าความสัมพันธ์ระหว่างคุณลักษณะทั้งหมดใน dataset กับ Label ของรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษในรูปแบบ correlation matrix ของการทดลองครั้งที่ 2 พบว่า feature ที่สำคัญ 3 อันดับแรกของการทำนายผลการเรียนวิชาคณิตศาสตร์4 ได้แก่ คณิตศาสตร์3 ที่ค่า 0.72 ผลการเรียนเฉลี่ยวิชาคณิตศาสตร์1-2 และผลการเรียนเฉลี่ยวิชาคณิตศาสตร์1-3 ที่ค่า 0.70 ผลการเรียนเฉลี่ยวิชาวิทยาศาสตร์1-3 ที่ค่า 0.61 สำหรับการทำนายผลการเรียนวิชาวิทยาศาสตร์4 ได้แก่ ผลการเรียนเฉลี่ยวิชาวิทยาศาสตร์1-2 และผลการเรียนเฉลี่ยวิชาวิทยาศาสตร์1-3 ที่ค่า 0.69 วิชาวิทยาศาสตร์3 ที่ค่า 0.68 ผลการเรียนเฉลี่ยวิชาคณิตศาสตร์1-3 ที่ค่า 0.67 สำหรับการทำนายผลการเรียนวิชาภาษาอังกฤษ4 ได้แก่ ผลการเรียนเฉลี่ยวิชาภาษาอังกฤษ1-2 ที่ค่า 0.60 วิชาภาษาอังกฤษ3 ที่ค่า 0.59 และผลการเรียนเฉลี่ยวิชาภาษาอังกฤษ1-3 ที่ค่า 0.58



ภาพประกอบ 16 แสดงตัวอย่างการหาค่าความสัมพันธ์(Correlation)
ระหว่าง Feature สำหรับการทดลองครั้งที่ 3

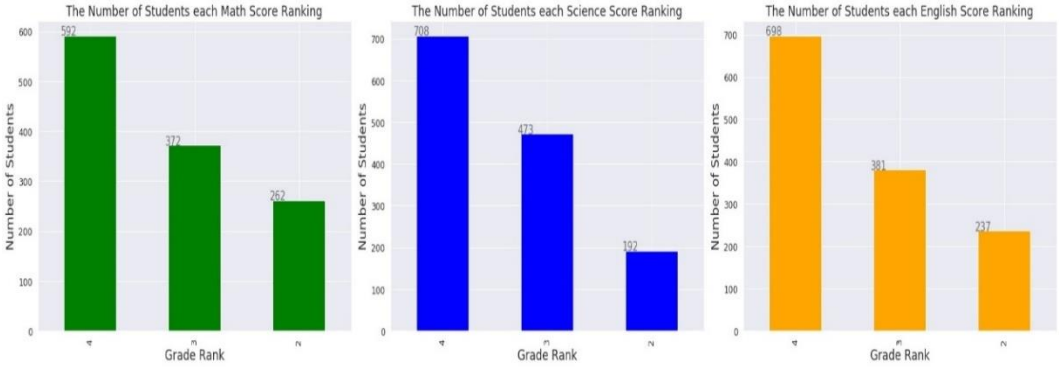
จากรูปภาพที่ 16 แสดงการหาค่าความสัมพันธ์ระหว่างคุณลักษณะทั้งหมดใน dataset กับ Label ของรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษในรูปแบบ correlation matrix ของการทดลองครั้งที่ 3 พบว่า feature ที่สำคัญ 3 อันดับแรกของการทำนายผลการเรียนวิชาคณิตศาสตร์5 ได้แก่ ผลการเรียนเฉลี่ยวิชาคณิตศาสตร์1-3 และผลการเรียนเฉลี่ยวิชาคณิตศาสตร์1-4 ที่ค่า 0.79 ผลการเรียนเฉลี่ยวิชาคณิตศาสตร์1-2 ที่ค่า 0.76 ผลการเรียนเฉลี่ยวิชาวิทยาศาสตร์1-3 และผลการเรียนเฉลี่ยวิชาวิทยาศาสตร์1-4 ที่ค่า 0.75 สำหรับการทำนายผลการเรียนวิชาวิทยาศาสตร์5 ได้แก่ ผลการเรียนเฉลี่ยวิชาคณิตศาสตร์1-4 ที่ค่า 0.69 ผลการเรียนเฉลี่ยวิชาคณิตศาสตร์1-3 ผลการเรียนเฉลี่ยวิทยาศาสตร์1-3 และผลการเรียนเฉลี่ยวิทยาศาสตร์1-4 ที่ค่า 0.68 สำหรับการทำนายผลการเรียนวิชาภาษาอังกฤษ5 ได้แก่ ผลการเรียนเฉลี่ยวิชาภาษาอังกฤษ1-3 และผลการเรียนเฉลี่ยวิชาภาษาอังกฤษ1-4 ที่ค่า 0.70 ผลการเรียนเฉลี่ยวิชาภาษาอังกฤษ1-2 ที่ค่า 0.67 ผลการเรียนเฉลี่ยวิชาคณิตศาสตร์1-3 ผลการเรียนเฉลี่ยวิชาคณิตศาสตร์1-4 และผลการเรียนเฉลี่ยวิชาวิทยาศาสตร์ ที่ค่า 0.66



ภาพประกอบ 17 แสดงตัวอย่างการหาค่าความสัมพันธ์(Correlation)ระหว่าง Feature สำหรับการทดลองครั้งที่ 4

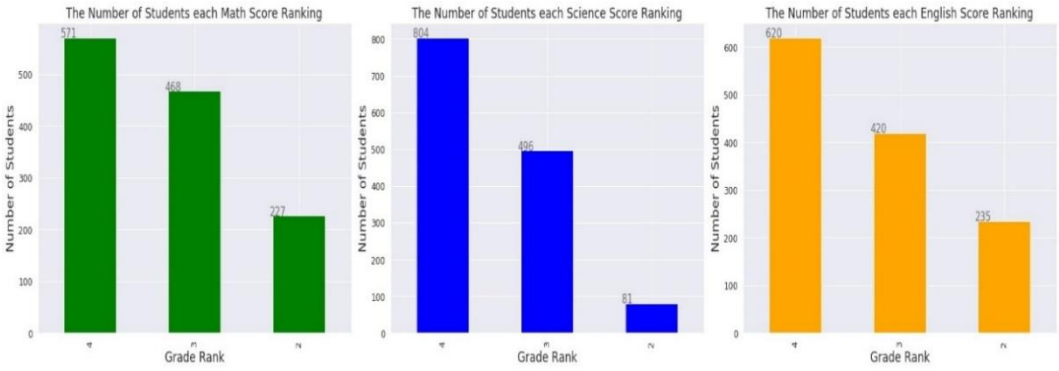
จากรูปภาพที่ 17 แสดงการหาค่าความสัมพันธ์ระหว่างคุณลักษณะทั้งหมดใน dataset กับ Label ของรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษในรูปแบบ correlation matrix ของการทดลองครั้งที่ 4 พบว่า feature ที่สำคัญ 3 อันดับแรกของการทำนายผลการเรียนวิชาคณิตศาสตร์6 ได้แก่ ผลการเรียนเฉลี่ยวิชาคณิตศาสตร์1-4 และผลการเรียนเฉลี่ยวิชาคณิตศาสตร์1-5 ที่ค่า 0.82 ผลการเรียนเฉลี่ยวิชาคณิตศาสตร์1-2 และผลการเรียนเฉลี่ยวิชาคณิตศาสตร์1-3 ที่ค่า 0.81 และวิชาคณิตศาสตร5 ที่ค่า 0.76 สำหรับการทำนายผลการเรียนวิชาวิทยาศาสตร์6 ได้แก่ ผลการเรียนเฉลี่ยวิชาวิทยาศาสตร์1-4 ที่ค่า 0.79 ผลการเรียนเฉลี่ยวิชาวิทยาศาสตร์1-3 และผลการเรียนเฉลี่ยวิทยาศาสตร์1-5 ที่ค่า 0.78 ผลการเรียนเฉลี่ยคณิตศาสตร์1-4 และผลการเรียนเฉลี่ยคณิตศาสตร์1-5 ที่ค่า 0.76 สำหรับการทำนายผลการเรียนวิชาภาษาอังกฤษ6 ได้แก่ ภาษาอังกฤษ5 ผลการเรียนเฉลี่ยวิชาภาษาอังกฤษ1-2 ผลการเรียนเฉลี่ยวิชาภาษาอังกฤษ1-3 และผลการเรียนเฉลี่ยวิชาภาษาอังกฤษ1-4 ที่ค่า 0.77 ผลการเรียนเฉลี่ยวิชาวิทยาศาสตร์1-5 ที่ค่า 0.76 และผลการเรียนเฉลี่ยวิชาคณิต1-4 ที่ค่า 0.71

ทำการเลือกข้อมูลระดับผลการเรียน(Label) ตั้งแต่ระดับพอใช้ ดี และดีมาก รวม 3 ระดับ ของรายวิชาคณิตศาสตร์, วิทยาศาสตร์ และภาษาอังกฤษ เพื่อสำรวจจำนวนข้อมูลในแต่ละระดับ(Class) ในการทดลองทั้ง 4 ครั้ง



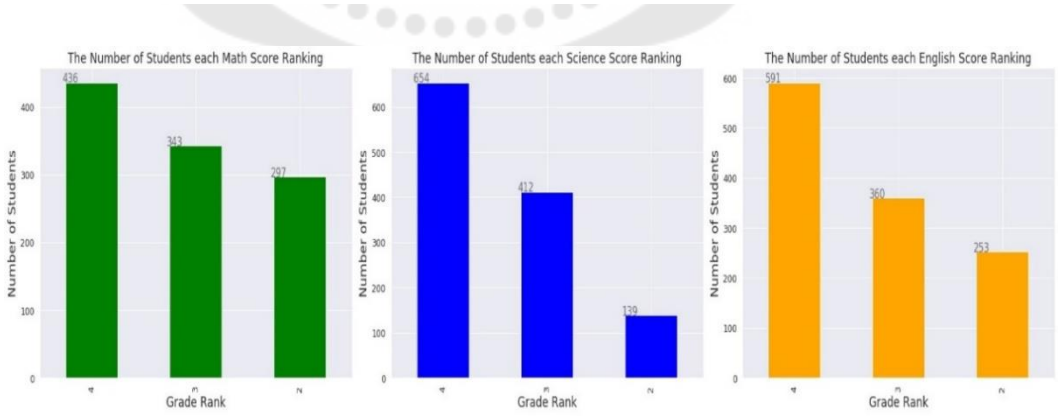
ภาพประกอบ 18 แสดงตัวอย่างจำนวนข้อมูลในแต่ละ class
ของผลการเรียนแต่ละรายวิชาสำหรับการทดลองครั้งที่ 1

จากรูปภาพที่ 18 แสดงตัวอย่างจำนวนข้อมูลผลการเรียนสำหรับการทดลองครั้งที่ 1 พบว่า รายวิชาคณิตศาสตร์3 มีนักเรียนที่ได้ผลการเรียนระดับ 2 หมายถึงระดับพอใช้ มีจำนวน 262 คน นักเรียนที่ได้ผลการเรียนระดับ 3 หมายถึงระดับ ดี มีจำนวน 372 คน และนักเรียนที่ได้ผลการเรียนระดับ 4 หมายถึงระดับดีมาก มีจำนวน 592 คน สำหรับรายวิชาวิทยาศาสตร์3 มีนักเรียนที่ได้ผลการเรียนระดับ 2 หมายถึงระดับพอใช้ มีจำนวน 192 คน นักเรียนที่ได้ผลการเรียนระดับ 3 หมายถึงระดับ ดี มีจำนวน 473 คน และนักเรียนที่ได้ผลการเรียนระดับ 4 หมายถึงระดับดีมาก มีจำนวน 708 คน และรายวิชาภาษาอังกฤษ3 มีนักเรียนที่ได้ผลการเรียนระดับ 2 หมายถึงระดับพอใช้ มีจำนวน 237 คน นักเรียนที่ได้ผลการเรียนระดับ 3 หมายถึงระดับ ดี มีจำนวน 381 คน และนักเรียนที่ได้ผลการเรียนระดับ 4 หมายถึงระดับดีมาก มีจำนวน 698 คน



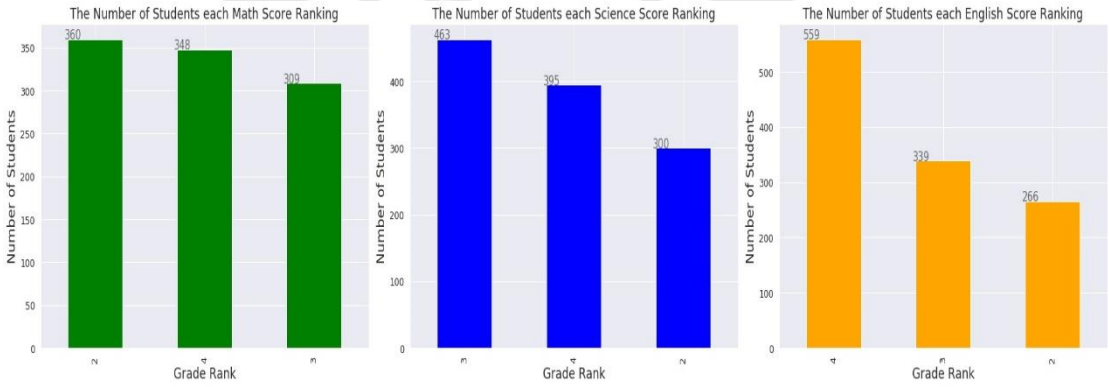
ภาพประกอบ 19 แสดงตัวอย่างจำนวนข้อมูลในแต่ละ class
ของผลการเรียนแต่ละรายวิชาสำหรับการทดลองครั้งที่ 2

จากรูปภาพที่ 19 แสดงตัวอย่างจำนวนข้อมูลผลการเรียนสำหรับการทดลองครั้งที่ 2 พบว่า รายวิชาคณิตศาสตร์4 มีนักเรียนที่ได้ผลการเรียนระดับ 2 หมายถึงระดับพอใช้ มีจำนวน 227 คน นักเรียนที่ได้ผลการเรียนระดับ 3 หมายถึงระดับ ดี มีจำนวน 468 คน และนักเรียนที่ได้ผลการเรียนระดับ 4 หมายถึงระดับดีมาก มีจำนวน 571 คน สำหรับรายวิชาวิทยาศาสตร์4 มีนักเรียนที่ได้ผลการเรียนระดับ 2 หมายถึงระดับพอใช้ มีจำนวน 81 คน นักเรียนที่ได้ผลการเรียนระดับ 3 หมายถึงระดับ ดี มีจำนวน 496 คน และนักเรียนที่ได้ผลการเรียนระดับ 4 หมายถึงระดับดีมาก มีจำนวน 804 คน และรายวิชาภาษาอังกฤษ4 มีนักเรียนที่ได้ผลการเรียนระดับ 2 หมายถึงระดับพอใช้ มีจำนวน 235 คน นักเรียนที่ได้ผลการเรียนระดับ 3 หมายถึงระดับ ดี มีจำนวน 420 คน และนักเรียนที่ได้ผลการเรียนระดับ 4 หมายถึงระดับดีมาก มีจำนวน 620 คน



ภาพประกอบ 20 แสดงตัวอย่างจำนวนข้อมูลในแต่ละ class
ของผลการเรียนแต่ละรายวิชาสำหรับการทดลองครั้งที่ 3

จากรูปภาพที่ 20 แสดงตัวอย่างจำนวนข้อมูลผลการเรียนสำหรับการทดลองครั้งที่ 3 พบว่า รายวิชาคณิตศาสตร์5 มีนักเรียนที่ได้ผลการเรียนระดับ 2 หมายถึงระดับพอใช้ มีจำนวน 297 คน นักเรียนที่ได้ผลการเรียนระดับ 3 หมายถึงระดับ ดี มีจำนวน 343 คน และนักเรียนที่ได้ผลการเรียนระดับ 4 หมายถึงระดับดีมาก มีจำนวน 436 คน สำหรับรายวิชาวิทยาศาสตร์5 มีนักเรียนที่ได้ผลการเรียนระดับ 2 หมายถึงระดับพอใช้ มีจำนวน 139 คน นักเรียนที่ได้ผลการเรียนระดับ 3 หมายถึงระดับ ดี มีจำนวน 412 คน และนักเรียนที่ได้ผลการเรียนระดับ 4 หมายถึงระดับดีมาก มีจำนวน 654 คน และรายวิชาภาษาอังกฤษ5 มีนักเรียนที่ได้ผลการเรียนระดับ 2 หมายถึงระดับพอใช้ มีจำนวน 253 คน นักเรียนที่ได้ผลการเรียนระดับ 3 หมายถึงระดับ ดี มีจำนวน 360 คน และนักเรียนที่ได้ผลการเรียนระดับ 4 หมายถึงระดับดีมาก มีจำนวน 591 คน



ภาพประกอบ 21 แสดงตัวอย่างจำนวนข้อมูลในแต่ละ class
ของผลการเรียนแต่ละรายวิชาสำหรับการทดลองครั้งที่ 4

จากรูปภาพที่ 21 แสดงตัวอย่างจำนวนข้อมูลผลการเรียนสำหรับการทดลองครั้งที่ 4 พบว่า รายวิชาคณิตศาสตร์6 มีนักเรียนที่ได้ผลการเรียนระดับ 2 หมายถึงระดับพอใช้ มีจำนวน 360 คน นักเรียนที่ได้ผลการเรียนระดับ 3 หมายถึงระดับ ดี มีจำนวน 309 คน และนักเรียนที่ได้ผลการเรียนระดับ 4 หมายถึงระดับดีมาก มีจำนวน 348 คน สำหรับรายวิชาวิทยาศาสตร์6 มีนักเรียนที่ได้ผลการเรียนระดับ 2 หมายถึงระดับพอใช้ มีจำนวน 300 คน นักเรียนที่ได้ผลการเรียนระดับ 3 หมายถึงระดับ ดี มีจำนวน 463 คน และนักเรียนที่ได้ผลการเรียนระดับ 4 หมายถึงระดับดีมาก มีจำนวน 395 คน และรายวิชาภาษาอังกฤษ6 มีนักเรียนที่ได้ผลการเรียนระดับ 2 หมายถึงระดับพอใช้ มีจำนวน 266 คน นักเรียนที่ได้ผลการเรียนระดับ 3 หมายถึงระดับ ดี มีจำนวน 339 คน และนักเรียนที่ได้ผลการเรียนระดับ 4 หมายถึงระดับดีมาก มีจำนวน 559 คน

สำหรับการทดลองทั้ง 4 ครั้งจะใช้เทคนิค Polynomial Feature สำหรับเพิ่ม feature และทำการ scale data เพื่อให้ค่าข้อมูลอยู่ในระดับเดียวกัน

math1^1	math2^1	math3^1	math4^1	math5^1	ave5_math^1	ave4_math^1	ave3_math^1	ave2_math^1	stdv5_math^1
0.585009	0.367291	-1.070943	-0.530363	0.104711	-0.140969	-0.326854	-0.517110	-0.195245	0.239959
0.585009	1.062677	-0.020238	0.520733	0.236775	0.524027	0.478421	0.262041	0.393578	-0.467901
0.666549	0.801907	0.430064	0.116465	-0.093384	0.416189	0.326072	0.154878	0.001029	-0.327475
0.095771	0.541138	-0.845792	-0.287802	-0.225448	-0.194888	-0.261562	-0.489714	-0.273755	-0.106806
0.095771	0.193445	-1.070943	-0.287802	-0.225448	-0.320698	-0.413911	-0.570288	-0.273755	-0.187049
...	...	...	...	...	...	...	...	...	...
-0.719626	-0.762710	-0.095289	0.682440	-0.753703	-0.392589	-0.283326	-0.114239	-0.116735	0.813124
-0.230388	-0.501941	-0.620641	-0.853777	-0.555608	-0.644209	-0.718610	-0.731436	-0.744813	-1.072590
-0.719626	-0.849634	0.054812	-0.449509	-1.017831	-0.698128	-0.653317	-0.543699	-0.823323	0.237093
-0.475007	-1.197326	-0.470540	-0.045242	-1.017831	-0.752046	-0.783903	-0.597684	-0.627049	0.205569
0.503470	-1.371173	-0.620641	-0.530363	-1.017831	-0.716100	-1.001545	-0.812010	-0.862578	0.684162

ภาพประกอบ 22 แสดงตัวอย่างข้อมูลหลังจากทำ Polynomial Feature และ scale data

จากรูปภาพที่ 22 แสดงตัวอย่างการเพิ่ม feature ด้วยเทคนิค polynomial feature หลังจากทำเทคนิคดังกล่าวจำนวน feature เพิ่มขึ้นดังนี้ ในการทดลองครั้งที่ 1 จำนวน feature เพิ่มจาก 12 feature เป็น 55 feature สำหรับการทดลองครั้งที่ 2 จำนวน feature เพิ่มจาก 18 feature เป็น 136 feature การทดลองครั้งที่ 3 จำนวน feature เพิ่มจาก 24 feature เป็น 253 feature และการทดลองครั้งที่ 4 จำนวน feature เพิ่มจาก 30 feature เป็น 406 feature จากนั้นทำการ scale ข้อมูลเพื่อให้ค่าข้อมูลทุกๆ feature อยู่ใน range เดียวกัน

สำหรับการทำ Feature Selection ใช้เทคนิค RFE เพื่อเลือก feature ที่มีความสำคัญกับข้อมูลคำตอบ(Label) ของแต่ละรายวิชาในการทดลองการทำนายทั้ง 4 ครั้ง

Math : Feature	importance
sci2^1	0.242197
ave_math^1	0.236201
math2^1	0.198011
math1^1 x sci1^1	0.076574
math2^1 x sci1^1	0.064509
math1^1 x eng1^1	0.048873
math2^1 x eng2^1	0.038801
math2^1 x eng1^1	0.037561
sci1^1 x eng2^1	0.034912
math1^1 x eng2^1	0.022361

Science : Feature	importance
ave_sci^1	0.254368
ave_math^1	0.174220
sci2^1	0.129362
sci1^1	0.125119
math2^1	0.090061
ave_math^1 x sci2^1	0.069111
sci1^1 x ave_eng^1	0.048567
math2^1 x sci1^1	0.044945
sci2^1 x eng2^1	0.035582
math1^1 x ave_eng^1	0.028666

English : Feature	importance
ave_sci^1	0.207398
ave_eng^1	0.193363
math2^1	0.162045
eng1^1	0.094188
eng2^1	0.094067
sci2^1 x eng2^1	0.069463
math1^1 x eng1^1	0.062163
math1^1 x math2^1	0.045842
math1^2	0.042403
math2^2	0.029069

ภาพประกอบ 23 แสดงตัวอย่าง feature สำคัญที่เลือกจากเทคนิค Feature Selection สำหรับการทดลองครั้งที่ 1

สำหรับการใช้เทคนิค Feature Selection ในการทดลองครั้งที่ 1 พบว่า feature ที่มีความสำคัญ 3 อันดับแรกสำหรับการทำนายผลการเรียนรายวิชาคณิตศาสตร์3 ได้แก่ sci2^1, ave_math^1 และ math2^1 สำหรับรายวิชาวิทยาศาสตร์3 ได้แก่ ave_sci^1, ave_math^1 และ sci2^1 สำหรับรายวิชาภาษาอังกฤษ3 ได้แก่ ave_sci^1, ave_eng^1 และmath2^1

Math : Feature	importance
math3^1	0.333066
mathave_3^1	0.172540
mathave_2^1	0.169401
engave_2^1	0.068389
math3^1 x sciave_2^1	0.065947
math2^1	0.064733
mathave_2^1 x eng3^1	0.053195
sciave_3^1 x engave_2^1	0.029814
sci1^1 x sci3^1	0.025944
sci1^1 x sci2^1	0.016970

Science : Feature	importance
mathave_3^1	0.295558
sci3^1	0.186014
sciave_3^1	0.152647
math2^1	0.119557
math1^1	0.060323
sci1^1	0.055771
math1^1 x sci1^1	0.038185
mathave_3^1 x eng2^1	0.035642
math1^1 x math3^1	0.029581
sciave_3^1 x eng2^1	0.026722

English : Feature	importance
eng3^1	0.276792
engave_3^1	0.202997
sci1^1	0.201596
sci3^1 x eng3^1	0.072749
math3^1 x eng1^1	0.056993
sci3^1 x engave_2^1	0.051958
sciave_2^1 x eng2^1	0.051854
eng1^1 x eng2^1	0.032054
sci1^1 x sci2^1	0.026751
sci1^1 x sci3^1	0.026256

ภาพประกอบ 24 แสดงตัวอย่าง feature สำคัญที่เลือกจากเทคนิค Feature Selection สำหรับการทดลองครั้งที่ 2

สำหรับการใช้เทคนิค Feature Selection ในการทดลองครั้งที่ 2 พบว่า feature ที่มีความสำคัญ 3 อันดับแรกสำหรับการทำนายผลการเรียนรายวิชาคณิตศาสตร์4 ได้แก่ math3^1,

mathave_3^1 และ mathave_2^1 สำหรับรายวิชาวิทยาศาสตร์4 ได้แก่ mathave_3^1, sci3^1 และ sciave_3^1 สำหรับรายวิชาภาษาอังกฤษ4 ได้แก่ eng3^1, engave_3^1 และsci1^1

Math : Feature	importance
math4_ave^1	0.244811
math3_ave^1	0.190293
math2_ave^1	0.153543
math3^1 x sci2_ave^1	0.097436
eng2_ave^1	0.086459
sci2_ave^1 x sci3_ave^1	0.069124
math3_ave^1 x sci1^1	0.056701
sci1^1	0.049226
math1^1 x math3_ave^1	0.031670
sci1^1 x sci3^1	0.020738

Science : Feature	importance
mathave_3^1	0.295558
sci3^1	0.186014
sciave_3^1	0.152647
math2^1	0.119557
math1^1	0.060323
sci1^1	0.055771
math1^1 x sci1^1	0.038185
mathave_3^1 x eng2^1	0.035642
math1^1 x math3^1	0.029581
sciave_3^1 x eng2^1	0.026722

English : Feature	importance
eng4_ave^1	0.243767
eng4^1	0.200610
math4^1	0.173029
sci2_ave^1	0.094891
math1^1 x eng4^1	0.083257
sci1^1	0.080742
sci2^1	0.039258
sci3^1 x eng4_ave^1	0.032532
eng3^1	0.032473
sci2^1 x eng3^1	0.019441

ภาพประกอบ 25 แสดงตัวอย่าง feature สำคัญที่เลือกจากเทคนิค Feature Selection สำหรับการทดลองครั้งที่ 3

สำหรับการใช้เทคนิค Feature Selection ในการทดลองครั้งที่ 3 พบว่า feature ที่มีความสำคัญ 3 อันดับแรกสำหรับการทำนายผลการเรียนรายวิชาคณิตศาสตร์5 ได้แก่ math4_eve^1, math3_ave^1 และ math2_ave^1 สำหรับรายวิชาวิทยาศาสตร์5 ได้แก่ mathave_3^1, sci3^1 และ sciave_3^1 สำหรับรายวิชาภาษาอังกฤษ5 ได้แก่ eng4_ave^1, eng4^1 และmath4^1

Math : Feature	importance
math5^1	0.288098
ave3_math^1	0.246419
ave4_math^1	0.195751
sci1^1	0.060788
ave2_math^1 x sci5^1	0.057131
math1^1 x math4^1	0.041114
math3^1 x sci1^1	0.038621
math2^1 x math3^1	0.034903
math3^1 x sci5^1	0.019001
sci1^1 x eng4^1	0.018174

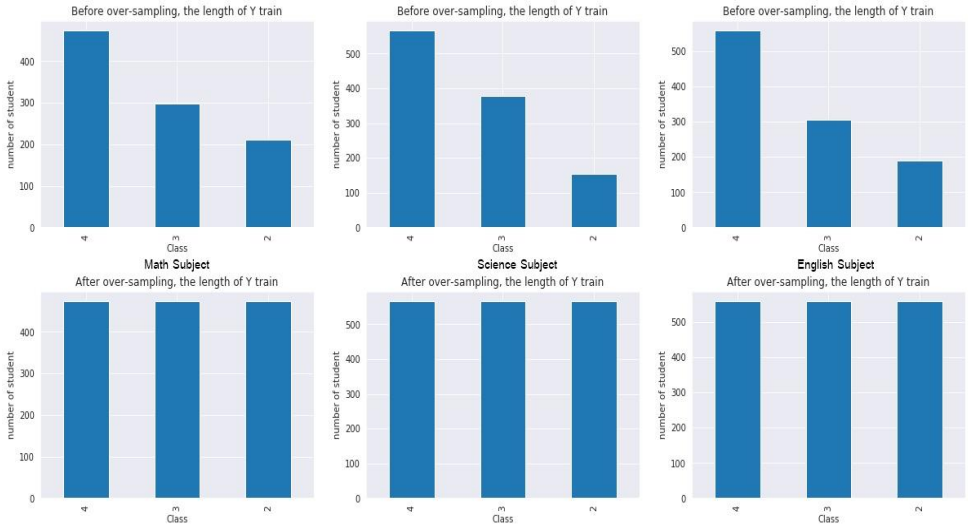
Science : Feature	importance
ave5_sci^1	0.369422
ave4_sci^1	0.337641
math5^1	0.109891
math5^1 x sci3^1	0.064122
math4^1 x sci3^1	0.043360
math1^1 x ave2_eng^1	0.021448
eng3^1 x eng4^1	0.019532
math1^1 x ave4_eng^1	0.015462
math1^1 x sci4^1	0.013460
eng3^2	0.005663

English : Feature	importance
ave5_eng^1	0.268476
ave2_eng^1	0.221928
eng5^1	0.213935
sci1^1	0.080756
eng5^1 x ave2_eng^1	0.074835
sci3^1 x eng4^1	0.054277
sci2^1 x ave2_sci^1	0.026666
math1^1 x sci1^1	0.023417
sci1^1 x ave3_sci^1	0.018173
sci1^1 x ave2_sci^1	0.017538

ภาพประกอบ 26 แสดงตัวอย่าง feature สำคัญที่เลือกจากเทคนิค Feature Selection สำหรับการทดลองครั้งที่ 4

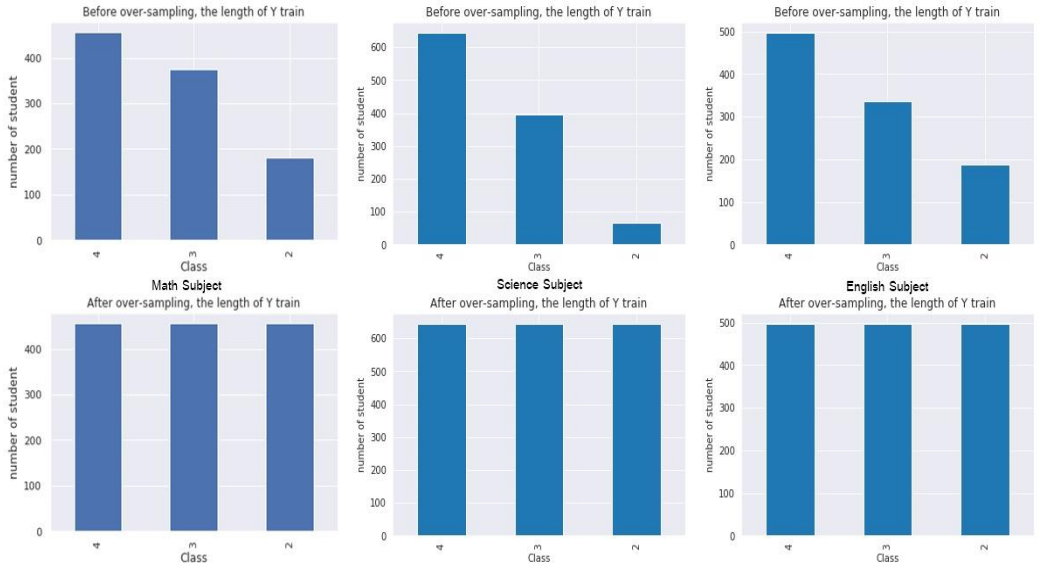
สำหรับการใช้เทคนิค Feature Selection ในการทดลองครั้งที่ 4 พบว่า feature ที่มีความสำคัญ 3 อันดับแรกสำหรับการทำนายผลการเรียนรายวิชาคณิตศาสตร์6 ได้แก่ math5^1, ave3_math^1 และ ave4_math^1 สำหรับรายวิชาวิทยาศาสตร์6 ได้แก่ ave5_sci^1, ave4_sci^1 และ math5^1 สำหรับรายวิชาภาษาอังกฤษ6 ได้แก่ ave5_eng^1, ave2_eng^1 และeng5^1

ผลจากการสำรวจจำนวนข้อมูลคำตอบ(Label) นำมาซึ่งการเพิ่มข้อมูลในแต่ละคลาสให้มีจำนวนเท่าๆกัน โดยใช้เทคนิค SMOTE จัดการปัญหา imbalance data ด้วยการเพิ่มจำนวนข้อมูลใน Class ที่มีจำนวนข้อมูลน้อยกว่า (Minority Class) ให้มีจำนวนข้อมูลเท่ากับClass ที่มีจำนวนข้อมูลมากที่สุด (Majority Class)



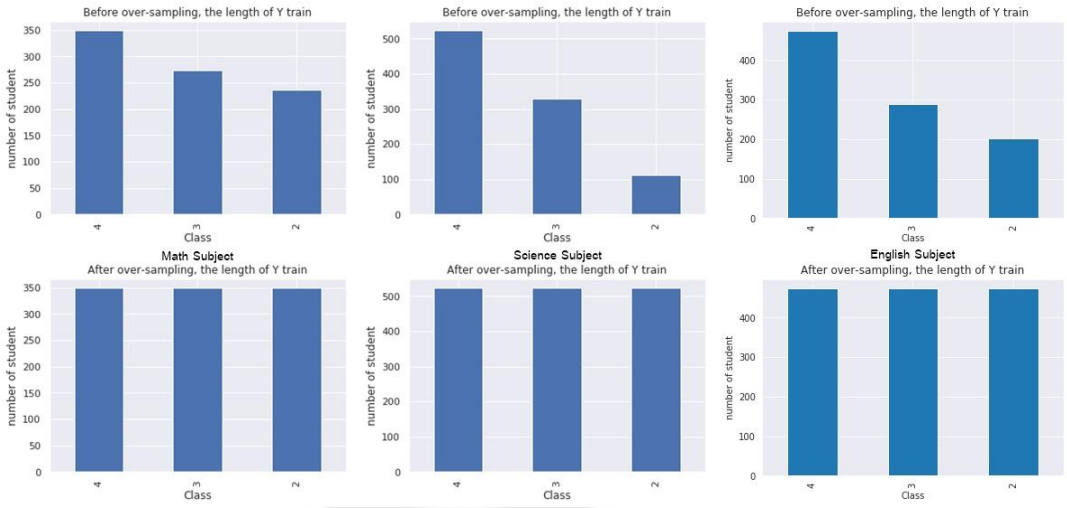
ภาพประกอบ 27 แสดงผลการทำ over sampling ด้วยเทคนิค SMOTE สำหรับการทดลองครั้งที่ 1

ในการทดลองที่ 1 มีจำนวนข้อมูลคำตอบในรายวิชาคณิตศาสตร์3ของคลาสดีมาก ดี และพอใช้ที่ 473, 297 และ 210 ตามลำดับ หลังจากทำการเพิ่มจำนวนข้อมูลด้วยเทคนิค SMOTE แล้วส่งผลให้จำนวนข้อมูลทั้ง 3 คลาสเพิ่มเป็น 473 เท่าๆกัน ในรายวิชาวิทยาศาสตร์3ที่คลาสดีมาก ดี และพอใช้ที่จำนวน 566, 378 และ 154 ตามลำดับ หลังจากทำการเพิ่มจำนวนข้อมูลด้วยเทคนิค SMOTE แล้วส่งผลให้จำนวนข้อมูลทั้ง 3 คลาสเพิ่มเป็น 566 เท่าๆกัน และในรายวิชาภาษาอังกฤษ3ที่คลาสดีมาก ดี และพอใช้ที่จำนวน 558, 305 และ 189 ตามลำดับ หลังจากทำการเพิ่มจำนวนข้อมูลด้วยเทคนิค SMOTE แล้วส่งผลให้จำนวนข้อมูลทั้ง 3 คลาสเพิ่มเป็น 558 เท่าๆกัน



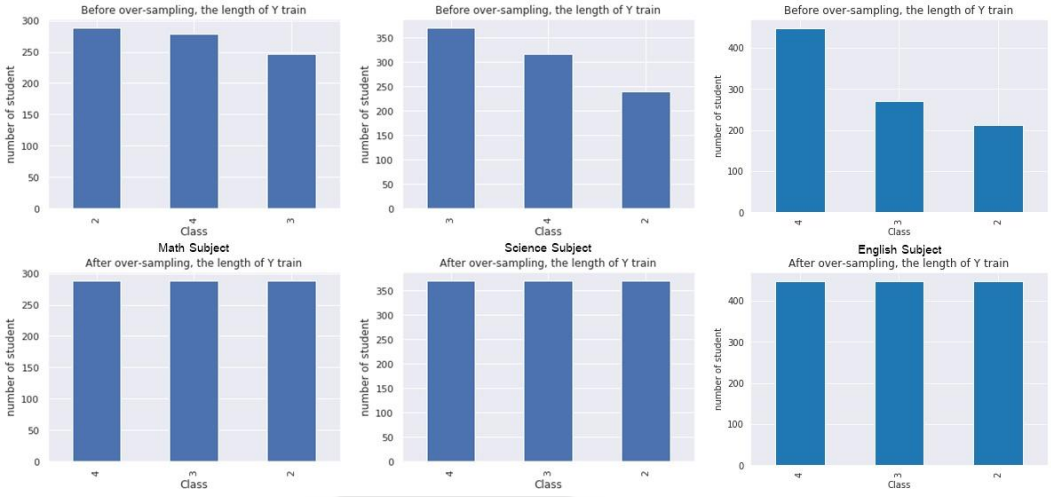
ภาพประกอบ 28 แสดงผลการทำ over sampling ด้วยเทคนิค SMOTE สำหรับการทดลองครั้งที่ 2

ในการทดลองที่ 2 มีจำนวนข้อมูลคำตอบในรายวิชาคณิตศาสตร์4ของคลาสดีมาก ดี และพอใช้ที่ 456, 374 และ 182 ตามลำดับ หลังจากทำการเพิ่มจำนวนข้อมูลด้วยเทคนิค SMOTE แล้วส่งผลให้จำนวนข้อมูลทั้ง 3 คลาสเพิ่มเป็น 456 เท่าๆกัน ในรายวิชาวิทยาศาสตร์4ที่คลาสดีมาก ดี และพอใช้ที่จำนวน 643, 396 และ 65 ตามลำดับ หลังจากทำการเพิ่มจำนวนข้อมูลด้วยเทคนิค SMOTE แล้วส่งผลให้จำนวนข้อมูลทั้ง 3 คลาสเพิ่มเป็น 643 เท่าๆกัน และในรายวิชาภาษาอังกฤษ4ที่คลาสดีมาก ดี และพอใช้ที่จำนวน 496, 336 และ 188 ตามลำดับ หลังจากทำการเพิ่มจำนวนข้อมูลด้วยเทคนิค SMOTE แล้วส่งผลให้จำนวนข้อมูลทั้ง 3 คลาสเพิ่มเป็น 496เท่าๆกัน



ภาพประกอบ 29 แสดงผลการทำ over sampling ด้วยเทคนิค SMOTE
สำหรับการทดลองครั้งที่ 3

ในการทดลองที่ 3 มีจำนวนข้อมูลคำตอบในรายวิชาคณิตศาสตร์5ของคลาสดีมาก ดี และพอใช้ที่ 349, 274 และ 237 ตามลำดับ หลังจากทำการเพิ่มจำนวนข้อมูลด้วยเทคนิค SMOTE แล้วส่งผลให้จำนวนข้อมูลทั้ง 3 คลาสเพิ่มเป็น 349 เท่าๆกัน ในรายวิชาวิทยาศาสตร์5ที่คลาสดีมาก ดี และพอใช้ที่จำนวน 523, 330 และ 111 ตามลำดับ หลังจากทำการเพิ่มจำนวนข้อมูลด้วยเทคนิค SMOTE แล้วส่งผลให้จำนวนข้อมูลทั้ง 3 คลาสเพิ่มเป็น 523 เท่าๆกัน และในรายวิชาภาษาอังกฤษ5ที่คลาสดีมาก ดี และพอใช้ที่จำนวน 473, 288 และ 202 ตามลำดับ หลังจากทำการเพิ่มจำนวนข้อมูลด้วยเทคนิค SMOTE แล้วส่งผลให้จำนวนข้อมูลทั้ง 3 คลาสเพิ่มเป็น 473เท่าๆกัน



ภาพประกอบ 30 แสดงผลการทำ over sampling ด้วยเทคนิค SMOTE สำหรับการทดลองครั้งที่ 4

ในการทดลองที่ 4 มีจำนวนข้อมูลคำตอบในรายวิชาคณิตศาสตร์6ของคลาสดีมาก ดี และพอใช้ที่ 288, 278 และ 247 ตามลำดับ หลังจากทำการเพิ่มจำนวนข้อมูลด้วยเทคนิค SMOTE แล้วส่งผลให้จำนวนข้อมูลทั้ง 3 คลาสเพิ่มเป็น 288 เท่าๆกัน ในรายวิชาวิทยาศาสตร์6ที่คลาสดีมาก ดี และพอใช้ที่จำนวน 370, 316 และ 240 ตามลำดับ หลังจากทำการเพิ่มจำนวนข้อมูลด้วยเทคนิค SMOTE แล้วส่งผลให้จำนวนข้อมูลทั้ง 3 คลาสเพิ่มเป็น 370 เท่าๆกัน และในรายวิชาภาษาอังกฤษ6ที่คลาสดีมาก ดี และพอใช้ที่จำนวน 447, 271 และ 213 ตามลำดับ หลังจากทำการเพิ่มจำนวนข้อมูลด้วยเทคนิค SMOTE แล้วส่งผลให้จำนวนข้อมูลทั้ง 3 คลาสเพิ่มเป็น 447เท่าๆกัน

การหาพารามิเตอร์ที่ดีที่สุดสำหรับการสร้างแบบจำลองการทำนายในการทดลองทั้ง 4 ครั้งของรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษ โดยใช้เทคนิค Grid Search และ 10 Folds Cross-Validation ในขั้นตอนการ train data เพื่อทำการ tuning hyper parameter และตรวจสอบประสิทธิภาพของแบบจำลองจากขั้นตอนดังกล่าวได้ผลลัพธ์ในแต่ละรายวิชาดังนี้

ตาราง 5 แสดงผล Hyper Parameter Tuning จากการทำ grid search โดยใช้เทคนิค Random Forest

การทดลอง	**รายวิชา**	max_leaf_nodes	max_depth	min_samples_split
ครั้งที่ 1	คณิตศาสตร์	19	271	2
	วิทยาศาสตร์	59	19	10
	ภาษาอังกฤษ	19	155	5
ครั้งที่ 2	คณิตศาสตร์	15	271	100
	วิทยาศาสตร์	59	19	10
	ภาษาอังกฤษ	24	97	5
ครั้งที่ 3	คณิตศาสตร์	25	6	2
	วิทยาศาสตร์	59	19	10
	ภาษาอังกฤษ	19	213	5
ครั้งที่ 4	คณิตศาสตร์	13	5	56
	วิทยาศาสตร์	20	26	50
	ภาษาอังกฤษ	19	21	5

จากตารางข้างต้นแสดงผลที่ได้จากการหาค่า hyper parameter จากอัลกอริทึม Random Forest โดยการ tuning ค่าในแต่ละ parameter ตามค่าเริ่มต้นที่กำหนดดังนี้

max_leaf_nodes กำหนดค่าระหว่าง 2, 3…60

max_depth กำหนดค่าระหว่าง 5, 6…300

min_samples_split กำหนดค่าระหว่าง 2, 3…100

ตาราง 6 แสดงผลการทำ Cross-Validation ของเทคนิค Random Forest

การทดลอง	รายวิชา	Accuracy
ครั้งที่ 1	คณิตศาสตร์	0.72
	วิทยาศาสตร์	0.70
	ภาษาอังกฤษ	0.62
ครั้งที่ 2	คณิตศาสตร์	0.70
	วิทยาศาสตร์	0.78
	ภาษาอังกฤษ	0.53
ครั้งที่ 3	คณิตศาสตร์	0.69
	วิทยาศาสตร์	0.77
	ภาษาอังกฤษ	0.67
ครั้งที่ 4	คณิตศาสตร์	0.70
	วิทยาศาสตร์	0.67
	ภาษาอังกฤษ	0.72

สำหรับผลลัพธ์การตรวจสอบประสิทธิภาพของแบบจำลองด้วยวิธีการทำ 10 Folds Cross-Validation โดยใช้ hyper parameter ที่ได้จากการทำ grid search ของอัลกอริทึม Random Forest พบว่า การทดลองที่ได้Accuracyดีที่สุดในรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษ คือการทดลองครั้งที่ 1, ครั้งที่ 2 และครั้งที่ 4 ตามลำดับ

ตาราง 7 แสดงผล Hyper Parameter Tuning จากการทำ grid search โดยใช้เทคนิค Logistic Regression โดยที่ Penalty = L2

การทดลอง	รายวิชา	C
ครั้งที่ 1	คณิตศาสตร์	0.01
	วิทยาศาสตร์	0.01
	ภาษาอังกฤษ	0.0001
ครั้งที่ 2	คณิตศาสตร์	0.00001
	วิทยาศาสตร์	0.01
	ภาษาอังกฤษ	0.00001
ครั้งที่ 3	คณิตศาสตร์	0.001
	วิทยาศาสตร์	0.01
	ภาษาอังกฤษ	0.001
ครั้งที่ 4	คณิตศาสตร์	0.1
	วิทยาศาสตร์	0.001
	ภาษาอังกฤษ	0.001

จากตารางข้างต้นแสดงผลที่ได้จากการหาค่า hyper parameter จากอัลกอริทึม Logistic Regression โดยการ tuning ค่าในแต่ละ parameter ตามค่าเริ่มต้นที่กำหนดดังนี้

C กำหนดค่าระหว่าง 0.00001, 0.0001…1

ตาราง 8 แสดงผลการทำ Cross-Validation ของเทคนิค Logistic Regression

การทดลอง	รายวิชา	Accuracy
ครั้งที่ 1	คณิตศาสตร์	0.70
	วิทยาศาสตร์	0.72
	ภาษาอังกฤษ	0.62
ครั้งที่ 2	คณิตศาสตร์	0.65
	วิทยาศาสตร์	0.76
	ภาษาอังกฤษ	0.57
ครั้งที่ 3	คณิตศาสตร์	0.65
	วิทยาศาสตร์	0.77
	ภาษาอังกฤษ	0.69
ครั้งที่ 4	คณิตศาสตร์	0.70
	วิทยาศาสตร์	0.67
	ภาษาอังกฤษ	0.73

สำหรับผลลัพธ์การตรวจสอบประสิทธิภาพของแบบจำลองด้วยวิธีการทำ 10 Folds Cross-Validation โดยใช้ hyper parameter ที่ได้จากการทำ grid search ของอัลกอริทึม Logistic Regression พบว่า การทดลองที่ได้Accuracyดีที่สุดในรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษ คือการทดลองครั้งที่ 1, ครั้งที่ 3 และครั้งที่ 4 ตามลำดับ

ตาราง 9 แสดงผล Hyper Parameter Tuning จากการทำ grid search โดยใช้เทคนิค K-Nearest Neighbors โดยที่ weight_options = distance

การทดลอง	รายวิชา	leaf_size	k_neighbor	metric_option
ครั้งที่ 1	คณิตศาสตร์	3	50	euclidean
	วิทยาศาสตร์	3	4	euclidean
	ภาษาอังกฤษ	3	44	euclidean
ครั้งที่ 2	คณิตศาสตร์	3	50	euclidean
	วิทยาศาสตร์	3	4	euclidean
	ภาษาอังกฤษ	30	68	manhattan
ครั้งที่ 3	คณิตศาสตร์	3	10	euclidean
	วิทยาศาสตร์	3	4	euclidean
	ภาษาอังกฤษ	3	30	manhattan
ครั้งที่ 4	คณิตศาสตร์	3	27	manhattan
	วิทยาศาสตร์	3	90	manhattan
	ภาษาอังกฤษ	50	54	manhattan

จากตารางข้างต้นแสดงผลที่ได้จากการหาค่า hyper parameter จากอัลกอริทึม K-Nearest Neighbors โดยการ tuning ค่าในแต่ละ parameter ตามค่าเริ่มต้นที่กำหนดดังนี้

leaf_size กำหนดค่าระหว่าง 3, 4…50

k_neighbor กำหนดค่าระหว่าง 3, 4…100

metric_option กำหนดค่า manhattan, euclidean และ minkowski

ตาราง 10 แสดงผลการทำ Cross-Validation ของเทคนิค K-Nearest Neighbors

การทดลอง	รายวิชา	Accuracy
ครั้งที่ 1	คณิตศาสตร์	0.70
	วิทยาศาสตร์	0.70
	ภาษาอังกฤษ	0.61
ครั้งที่ 2	คณิตศาสตร์	0.72
	วิทยาศาสตร์	0.70
	ภาษาอังกฤษ	0.53
ครั้งที่ 3	คณิตศาสตร์	0.67
	วิทยาศาสตร์	0.70
	ภาษาอังกฤษ	0.68
ครั้งที่ 4	คณิตศาสตร์	0.69
	วิทยาศาสตร์	0.65
	ภาษาอังกฤษ	0.72

สำหรับผลลัพธ์การตรวจสอบประสิทธิภาพของแบบจำลองด้วยวิธีการทำ 10 Folds Cross-Validation โดยใช้ hyper parameter ที่ได้จากการทำ grid search ของอัลกอริทึม K-Nearest Neighbors พบว่า การทดลองที่ได้Accuracyดีที่สุดในรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษ คือการทดลองครั้งที่ 2, ครั้งที่ 1 และครั้งที่ 4 ตามลำดับ

ตาราง 11 แสดงผล Hyper Parameter Tuning จากการทำ grid search โดยใช้เทคนิค Support Vector Machine

การทดลอง	รายวิชา	kernel	gamma	C
ครั้งที่ 1	คณิตศาสตร์	rbf	0.0001	1000
	วิทยาศาสตร์	rbf	0.00001	1000
	ภาษาอังกฤษ	rbf	0.00001	1000
ครั้งที่ 2	คณิตศาสตร์	rbf	0.01	10
	วิทยาศาสตร์	rbf	0.00001	1000
	ภาษาอังกฤษ	rbf	0.0001	10
ครั้งที่ 3	คณิตศาสตร์	sigmoid	0.001	20
	วิทยาศาสตร์	rbf	0.00001	1000
	ภาษาอังกฤษ	rbf	0.001	10
ครั้งที่ 4	คณิตศาสตร์	sigmoid	0.001	100
	วิทยาศาสตร์	rbf	0.0001	100
	ภาษาอังกฤษ	rbf	0.0001	100

จากตารางข้างต้นแสดงผลที่ได้จากการหาค่า hyper parameter จากอัลกอริทึม Support Vector Machine โดยการ tuning ค่าในแต่ละ parameter ตามค่าเริ่มต้นที่กำหนดดังนี้

Kernel กำหนดค่า rbf, sigmoid และ linear

gamma กำหนดค่าระหว่าง 0.00001, 0.0001…1

C กำหนดค่าระหว่าง 10, 20…1000

ตาราง 12 แสดงผลการทำ Cross-Validation ของเทคนิค Support Vector Machine

การทดลอง	รายวิชา	Accuracy
ครั้งที่ 1	คณิตศาสตร์	0.67
	วิทยาศาสตร์	0.70
	ภาษาอังกฤษ	0.59
ครั้งที่ 2	คณิตศาสตร์	0.71
	วิทยาศาสตร์	0.74
	ภาษาอังกฤษ	0.57
ครั้งที่ 3	คณิตศาสตร์	0.64
	วิทยาศาสตร์	0.74
	ภาษาอังกฤษ	0.67
ครั้งที่ 4	คณิตศาสตร์	0.70
	วิทยาศาสตร์	0.65
	ภาษาอังกฤษ	0.72

สำหรับผลลัพธ์การตรวจสอบประสิทธิภาพของแบบจำลองด้วยวิธีการทำ 10 Folds Cross-Validation โดยใช้ hyper parameter ที่ได้จากการทำ grid search ของอัลกอริทึม Support Vector Machine พบว่า การทดลองที่ได้Accuracyดีที่สุดในรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษ คือการทดลองครั้งที่ 2, ครั้งที่ 2 และครั้งที่ 4 ตามลำดับ

ตาราง 13 แสดงผล Hyper Parameter Tuning จากการทำ grid search โดยใช้เทคนิค XGBoost

การทดลอง	รายวิชา	Learning rate	Max depth	N estimators	Colsample bytree	subsample
ครั้งที่ 1	คณิตศาสตร์	0.001	4	79	0.9	0.9
	วิทยาศาสตร์	0.00001	14	200	1	0.9
	ภาษาอังกฤษ	0.00001	80	38	0.01	0.01
ครั้งที่ 2	คณิตศาสตร์	1	12	120	1	0.9
	วิทยาศาสตร์	0.01	5	180	1	0.9
	ภาษาอังกฤษ	0.001	20	128	0.9	0.9
ครั้งที่ 3	คณิตศาสตร์	0.01	25	31	0.9	0.1
	วิทยาศาสตร์	0.01	5	180	1	0.9
	ภาษาอังกฤษ	0.01	77	128	0.9	0.9
ครั้งที่ 4	คณิตศาสตร์	0.01	76	180	0.9	0.9
	วิทยาศาสตร์	0.0001	4	151	0.9	0.9
	ภาษาอังกฤษ	0.01	2	180	0.9	0.9

จากตารางข้างต้นแสดงผลที่ได้จากการหาค่า hyper parameter จากอัลกอริทึม XGBoost โดยการ tuning ค่าในแต่ละ parameter ตามค่าเริ่มต้นที่กำหนดดังนี้

Learning rate กำหนดค่าระหว่าง 0.00001, 0.0001…1

Max depth กำหนดค่าระหว่าง 2, 3…100

n_estimators กำหนดค่าระหว่าง 30, 31…200

Colsample by tree กำหนดค่าระหว่าง 0.01, 0.02…1

subsample กำหนดค่าระหว่าง 0.01, 0.02…1

ตาราง 14 แสดงผลการทำ Cross-Validation ของเทคนิค XGBoost

การทดลอง	รายวิชา	Accuracy
ครั้งที่ 1	คณิตศาสตร์	0.70
	วิทยาศาสตร์	0.72
	ภาษาอังกฤษ	0.60
ครั้งที่ 2	คณิตศาสตร์	0.69
	วิทยาศาสตร์	0.76
	ภาษาอังกฤษ	0.57
ครั้งที่ 3	คณิตศาสตร์	0.64
	วิทยาศาสตร์	0.75
	ภาษาอังกฤษ	0.64
ครั้งที่ 4	คณิตศาสตร์	0.67
	วิทยาศาสตร์	0.70
	ภาษาอังกฤษ	0.73

สำหรับผลลัพธ์การตรวจสอบประสิทธิภาพของแบบจำลองด้วยวิธีการทำ 10 Folds Cross-Validation โดยใช้ hyper parameter ที่ได้จากการทำ grid search ของอัลกอริทึม XGBoost พบว่า การทดลองที่ได้Accuracyดีที่สุดในรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษ คือการทดลองครั้งที่ 1, ครั้งที่ 2 และครั้งที่ 4 ตามลำดับ

ตาราง 15 แสดงการเปรียบเทียบผลการทำ Cross-Validation ของแต่ละเทคนิค

การทดลอง	รายวิชา	Accuracy				
		Random Forest	Logistic Regression	KNN	SVM	XGBoost
ครั้งที่ 1	คณิตศาสตร์	0.72	0.70	0.70	0.67	0.70
	วิทยาศาสตร์	0.70	0.72	0.70	0.70	0.72
	ภาษาอังกฤษ	0.62	0.62	0.61	0.59	0.60
ครั้งที่ 2	คณิตศาสตร์	0.70	0.65	0.72	0.71	0.69
	วิทยาศาสตร์	0.78	0.76	0.70	0.74	0.76
	ภาษาอังกฤษ	0.53	0.57	0.53	0.57	0.57
ครั้งที่ 3	คณิตศาสตร์	0.69	0.65	0.67	0.64	0.64
	วิทยาศาสตร์	0.77	0.77	0.70	0.74	0.75
	ภาษาอังกฤษ	0.67	0.69	0.68	0.67	0.64
ครั้งที่ 4	คณิตศาสตร์	0.70	0.70	0.69	0.70	0.67
	วิทยาศาสตร์	0.67	0.67	0.65	0.65	0.70
	ภาษาอังกฤษ	0.72	0.73	0.72	0.72	0.73

จากการเปรียบผลลัพธ์การตรวจสอบประสิทธิภาพจากการทำ 10 Fols Cross-Validate ด้วยค่า Hyperparameter ที่ได้จากการเทคนิค Grid Search พบว่าเทคนิคที่ให้ประสิทธิภาพดีที่สุดในรายวิชาคณิตศาสตร์จากการทดลองครั้งที่ 1 คือ Random Forest ที่ค่า Accuracy เท่ากับ 0.70, ครั้งที่ 2 คือ K-Nearest Neighbor ที่ค่า Accuracy เท่ากับ 0.72, ครั้งที่ 3 Random Forest ที่ค่า Accuracy เท่ากับ 0.69 และครั้งที่ 4 ได้แก่ Random Forest, Logistic Regression และ Support Vector Machine ที่ค่า Accuracy เท่ากับ 0.70 ในส่วนของเทคนิคที่ให้ประสิทธิภาพดีที่สุดในรายวิชาวิทยาศาสตร์จากการทดลองครั้งที่ 1 คือ Logistic Regression และ XGBoost ที่ค่า Accuracy เท่ากับ 0.72, ครั้งที่ 2 คือ Random Forest ที่ค่า Accuracy เท่ากับ 0.78, ครั้งที่ 3 คือ Random Forest, Logistic Regression ที่ค่า Accuracy เท่ากับ 0.77 และครั้งที่ 4 คือ XGBoost ที่ค่า Accuracy เท่ากับ 0.70 และในส่วนของเทคนิคที่ให้ประสิทธิภาพดีที่สุดในรายวิชาภาษาอังกฤษจากการทดลองครั้งที่ 1 คือ Random Forest ที่ค่า Accuracy เท่ากับ 0.62 ครั้งที่ 2

คือ Logistic Regression, Support Vector Machine และ XGBoost ที่ค่า Accuracy เท่ากับ 0.57 ครั้งที่ 3 คือ Logistic Regression ค่า Accuracy เท่ากับ 0.69 และครั้งที่ 4 คือ Logistic Regression และ XGBoost ที่ค่า Accuracy เท่ากับ 0.73

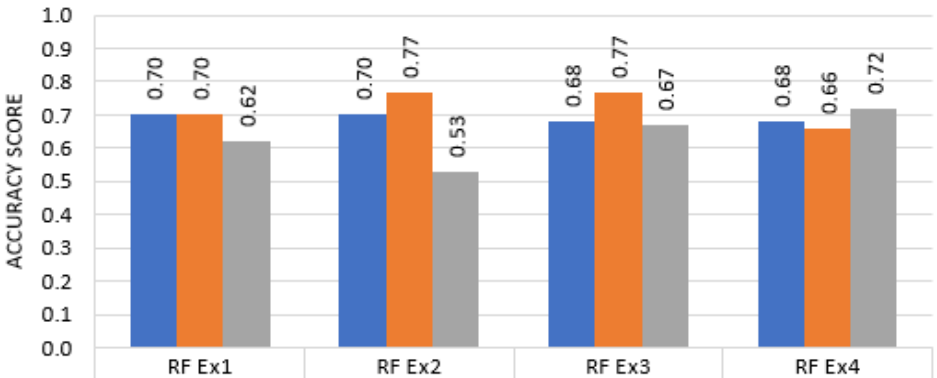
บทที่ 4
ผลการศึกษา

ในการวิจัยการทำนายผลการเรียนของนักเรียนในรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษ โดยใช้ข้อมูลคะแนนรายวิชาจากโรงเรียนระดับมัธยมศึกษาแห่งหนึ่งในจังหวัดสุพรรณบุรี โดยใช้เทคนิคการเรียนรู้ของเครื่อง ผู้วิจัยได้ดำเนินการวิจัยโดยศึกษาตามขั้นตอนต่างๆ ตลอดจนวัดประสิทธิภาพ เพื่อให้บรรลุจุดประสงค์ของการวิจัยที่ได้กำหนดไว้ได้ ดังนี้

1. ผลลัพธ์ของการสร้างแบบจำลอง Random Forest
2. ผลลัพธ์ของการสร้างแบบจำลอง Logistic Regression
3. ผลลัพธ์ของการสร้างแบบจำลอง K-Nearest Neighbor
4. ผลลัพธ์ของการสร้างแบบจำลอง Support Vector Machine
5. ผลลัพธ์ของการสร้างแบบจำลอง Extreme Gradient Boosting
6. ผลลัพธ์จากการเปรียบเทียบการทำนายผลการเรียนของนักเรียนของแบบจำลอง

ผลลัพธ์ของการสร้างแบบจำลอง Random Forest

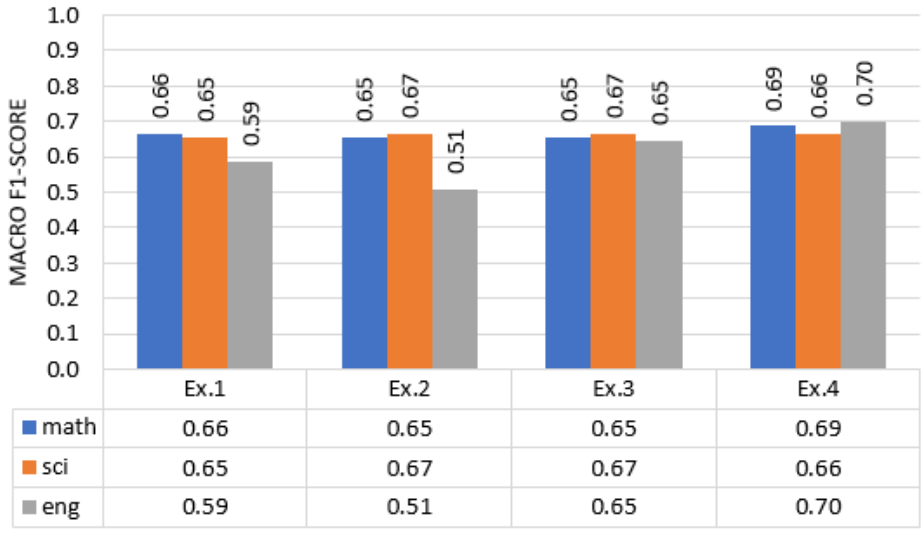
จากการทดลองเพื่อทำนายผลการเรียนรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษ โดยใช้แบบจำลอง Random Forest ทั้ง 4 ครั้ง ได้ผลลัพธ์ดังนี้



	RF Ex1	RF Ex2	RF Ex3	RF Ex4
math	0.70	0.70	0.68	0.68
sci	0.70	0.77	0.77	0.66
eng	0.62	0.53	0.67	0.72

ภาพประกอบ 31 แสดงผลการทำนายผลการเรียนของนักเรียนด้วยแบบจำลอง Random Forest

จากรูปภาพที่ 31 แสดงผลการทำนายผลการเรียนของนักเรียนด้วยแบบจำลอง Random Forest จากการทดลองทั้ง 4 ครั้ง พบว่าในการทดลองครั้งที่ 1 แบบจำลองได้ค่า accuracy ในรายวิชาคณิตศาสตร์ที่ 0.70 รายวิชาวิทยาศาสตร์ 0.70 และรายวิชาภาษาอังกฤษ 0.62 ต่อมาในการทดลองครั้งที่ 2 แบบจำลองได้ค่า accuracy ในรายวิชาคณิตศาสตร์ที่ 0.70 รายวิชาวิทยาศาสตร์ 0.77 และรายวิชาภาษาอังกฤษ 0.53 สำหรับการทดลองครั้งที่ 3 แบบจำลองได้ค่า accuracy ในรายวิชาคณิตศาสตร์ที่ 0.68 รายวิชาวิทยาศาสตร์ 0.77 และรายวิชาภาษาอังกฤษ 0.67 และในการทดลองครั้งที่ 4 แบบจำลองได้ค่า accuracy ในรายวิชาคณิตศาสตร์ที่ 0.68 รายวิชาวิทยาศาสตร์ 0.66 และรายวิชาภาษาอังกฤษ 0.72 ตามลำดับ



	Ex.1	Ex.2	Ex.3	Ex.4
math	0.66	0.65	0.65	0.69
sci	0.65	0.67	0.67	0.66
eng	0.59	0.51	0.65	0.70

ภาพประกอบ 32 แสดงค่า Macro F1-Score ในการทำนายผลการเรียนของนักเรียนด้วยแบบจำลอง Random Forest

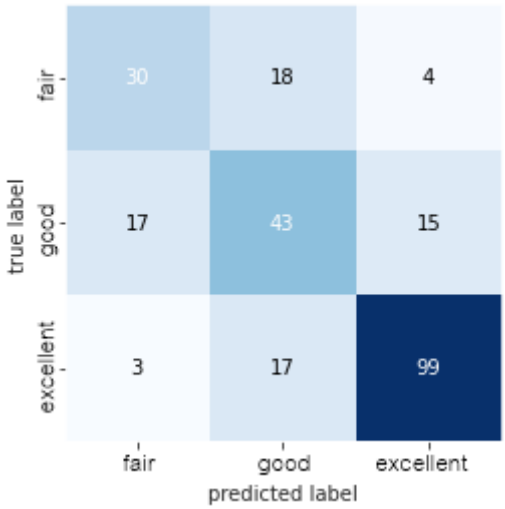
จากรูปภาพที่ 32 แสดงผล Macro F1-Score จากการทำนายผลการเรียนของนักเรียนด้วยแบบจำลอง Random Forest จากการทดลองทั้ง 4 ครั้ง พบว่าในการทดลองครั้งที่ 1 แบบจำลองได้ค่า Macro F1-Score รายวิชาคณิตศาสตร์ รายวิชาวิทยาศาสตร์ และรายวิชาภาษาอังกฤษ ที่ 0.66, 0.65 และ 0.59 ตามลำดับ ในการทดลองครั้งที่ 2 ที่ค่า 0.65, 0.67 และ 0.51 ตามลำดับ ในการทดลองครั้งที่ 3 ที่ค่า 0.65, 0.67 และ 0.65 ตามลำดับและในการทดลองครั้งที่ 4 ที่ค่า 0.69, 0.66 และ 0.70 ตามลำดับ

การทดลองเพื่อทำนายผลการเรียนทั้ง 4 ครั้งนั้นพบว่าแบบจำลอง Random Forest ให้ประสิทธิภาพการทำงานที่ดีที่สุดสำหรับการทำนายรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษ ในการทดลองการทำนายครั้งที่ 1, การทดลองการทำนายครั้งที่ 2 และการทดลองการทำนายครั้งที่ 4 ตามลำดับ ซึ่งสามารถแสดงรายละเอียดประสิทธิภาพได้ดังต่อไปนี้

ตาราง 16 แสดงรายละเอียดประสิทธิภาพการทำงานของแบบจำลอง Random Forest ในการทดลองการทำนายที่ 1 ในรายวิชาคณิตศาสตร์

คลาส	Precision	Recall	F1-Score
พอใช้	0.60	0.58	0.59
ดี	0.55	0.57	0.56
ดีมาก	0.84	0.83	0.84

จากตารางพบว่าแบบจำลอง Random Forest มีประสิทธิภาพการทำนายผลการเรียนรายวิชาคณิตศาสตร์ที่ดีที่สุดในคลาสดีมากที่ค่า precision 84% recall 83% และf1 84% รองลงมาคือคลาสพอใช้ที่ค่า precision 60% recall 58% และf1 59% และลำดับสุดท้ายคือคลาสดีที่ค่า precision 55% recall 57% และf1 56% ตามลำดับ



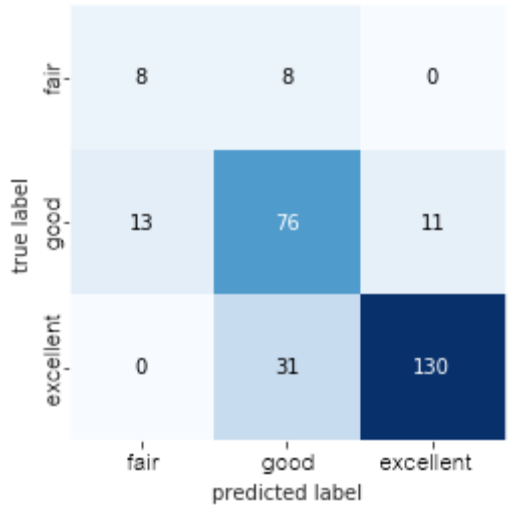
ภาพประกอบ 33 แสดงค่า confusion matrix ของแบบ Random Forest ในการทดลองการทำนายที่ 1 ในรายวิชาคณิตศาสตร์

จากค่า confusion matrix พบว่าประสิทธิภาพการทำนายของแบบจำลอง Random Forest ในรายวิชาคณิตศาสตร์สามารถทำนายได้ถูกต้องในคลาสพอใช้ 30 จาก 52 คน ทำนายได้ถูกต้องในคลาสดี 43 จาก 75 คน และทำนายได้ถูกต้องในคลาสดีมาก 99 จาก 119 คน

ตาราง 17 แสดงรายละเอียดประสิทธิภาพการทำงานของแบบจำลอง Random Forest ในการทดลองการทำนายที่ 2 ในรายวิชาวิทยาศาสตร์

คลาส	Precision	Recall	F1-Score
พอใช้	0.38	0.50	0.43
ดี	0.66	0.76	0.71
ดีมาก	0.92	0.81	0.86

จากตารางพบว่าแบบจำลอง Random Forest มีประสิทธิภาพการทำนายผลการเรียนรายวิชาวิทยาศาสตร์ที่ดีที่สุดในคลาสดีมากที่ค่า precision 92% recall 81% และf1 86% รองลงมาคือคลาสดีที่ค่า precision 66% recall 76% และf1 71% และลำดับสุดท้ายคือคลาสพอใช้ที่ค่า precision 38% recall 50% และf1 43% ตามลำดับ



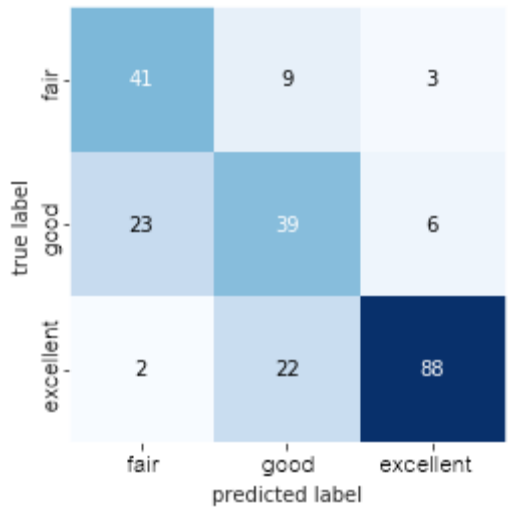
ภาพประกอบ 34 แสดงค่า confusion matrix ของแบบ Random Forest ในการทดลองการทำนายที่ 2 ในรายวิชาวิทยาศาสตร์

จากค่า confusion matrix พบว่าประสิทธิภาพการทำนายของแบบจำลอง Random Forest ในรายวิชาวิทยาศาสตร์สามารถทำนายได้ถูกต้องในคลาสพอใช้ 8 จาก 16 คน ทำนายได้ถูกต้องในคลาสดี 76 จาก 100 คน และทำนายได้ถูกต้องในคลาสดีมาก 130 จาก 161 คน

ตาราง 18 แสดงรายละเอียดประสิทธิภาพการทำงานของแบบจำลอง Random Forest ในการทดลองการทำนายที่ 4 ในรายวิชาภาษาอังกฤษ

คลาส	Precision	Recall	F1-Score
พอใช้	0.62	0.77	0.69
ดี	0.56	0.57	0.57
ดีมาก	0.91	0.79	0.84

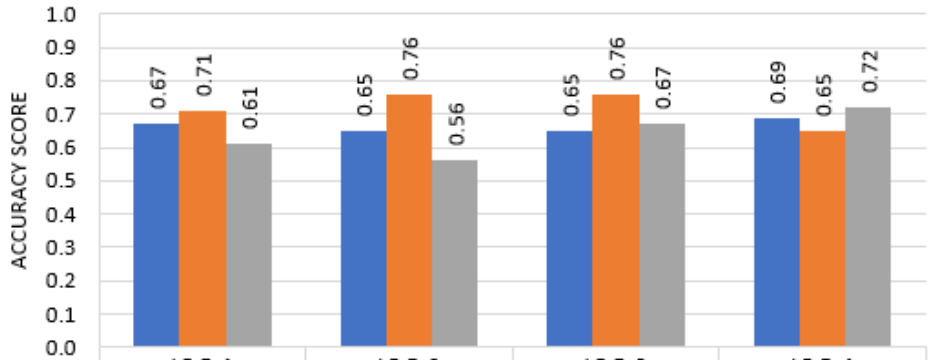
จากตารางพบว่าแบบจำลอง Random Forest มีประสิทธิภาพการทำนายผลการเรียนรายวิชาภาษาอังกฤษที่ดีที่สุดในคลาสดีมากที่ค่า precision 91% recall 79% และf1 84% รองลงมาคือคลาสพอใช้ที่ค่า precision 62% recall 77% และf1 69% และลำดับสุดท้ายคือคลาสดีที่ค่า precision 56% recall 57% และf1 57% ตามลำดับ



ภาพประกอบ 35 แสดงค่า confusion matrix ของแบบ Random Forest ในการทดลองการทำนายที่ 4 ในรายวิชาภาษาอังกฤษ

จากค่า confusion matrix พบว่าประสิทธิภาพการทำนายของแบบจำลอง Random Forest ในรายวิชาภาษาอังกฤษสามารถทำนายได้ถูกต้องในคลาสพอใช้ 41 จาก 53 คน ทำนายได้ถูกต้องในคลาสดี 39 จาก 68 คน และทำนายได้ถูกต้องในคลาสดีมาก 88 จาก 112 คน

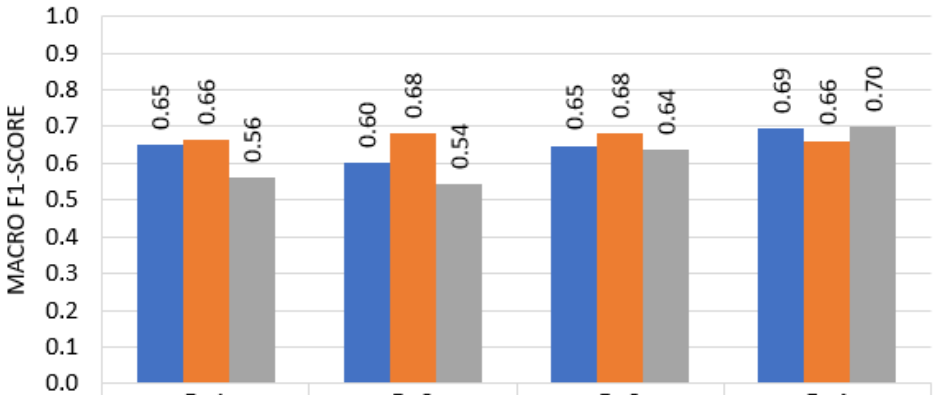
ผลลัพธ์ของการสร้างแบบจำลอง Logistic Regression



	LR Ex1	LR Ex2	LR Ex3	LR Ex4
math	0.67	0.65	0.65	0.69
sci	0.71	0.76	0.76	0.65
eng	0.61	0.56	0.67	0.72

ภาพประกอบ 36 แสดงผลการทำนายผลการเรียนของนักเรียน
ด้วยแบบจำลอง Logistic Regression

จากรูปภาพที่ 36 แสดงผลการทำนายผลการเรียนของนักเรียนด้วยแบบจำลอง Logistic Regression จากการทดลองทั้ง 4 ครั้ง พบว่าในการทดลองครั้งที่ 1 แบบจำลองได้ค่า accuracy ในรายวิชาคณิตศาสตร์ที่ 0.67 รายวิชาวิทยาศาสตร์ 0.71 และรายวิชาภาษาอังกฤษ 0.61 ต่อมาในการทดลองครั้งที่ 2 แบบจำลองได้ค่า accuracy ในรายวิชาคณิตศาสตร์ที่ 0.65 รายวิชาวิทยาศาสตร์ 0.76 และรายวิชาภาษาอังกฤษ 0.56 สำหรับการทดลองครั้งที่ 3 แบบจำลองได้ค่า accuracy ในรายวิชาคณิตศาสตร์ที่ 0.65 รายวิชาวิทยาศาสตร์ 0.76 และรายวิชาภาษาอังกฤษ 0.67 และในการทดลองครั้งที่ 4 แบบจำลองได้ค่า accuracy ในรายวิชาคณิตศาสตร์ที่ 0.69 รายวิชาวิทยาศาสตร์ 0.65 และรายวิชาภาษาอังกฤษ 0.72 ตามลำดับ



	Ex.1	Ex.2	Ex.3	Ex.4
math	0.65	0.60	0.65	0.69
sci	0.66	0.68	0.68	0.66
eng	0.56	0.54	0.64	0.70

ภาพประกอบ 37 แสดงค่า Macro F1-Score ในการทำนายผลการเรียนของนักเรียนด้วยแบบจำลอง Logistic Regression

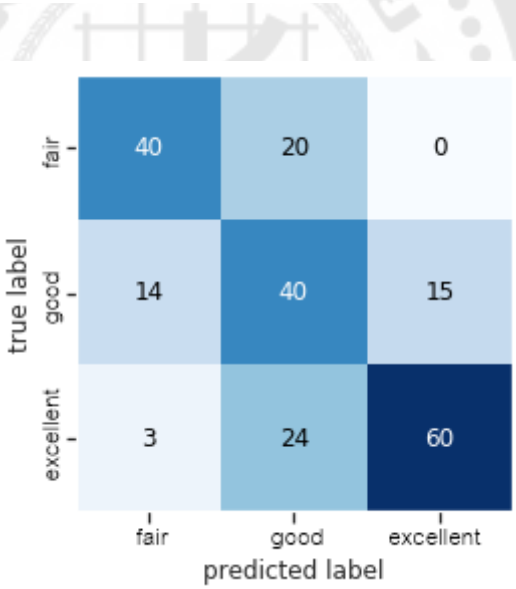
จากรูปภาพที่ 37 แสดงผล Macro F1-Score จากการทำนายผลการเรียนของนักเรียนด้วยแบบจำลอง Logistic Regression จากการทดลองทั้ง 4 ครั้ง พบว่าในการทดลองครั้งที่ 1 แบบจำลองได้ค่า Macro F1-Score รายวิชาคณิตศาสตร์ รายวิชาวิทยาศาสตร์ และรายวิชาภาษาอังกฤษ ที่ 0.65, 0.66 และ0.56 ตามลำดับ ในการทดลองครั้งที่ 2 ที่ค่า 0.60, 0.68 และ 0.54 ตามลำดับ ในการทดลองครั้งที่ 3 ที่ค่า 0.65, 0.68 และ0.64 ตามลำดับและในการทดลองครั้งที่ 4 ที่ค่า 0.69, 0.66 และ0.70 ตามลำดับ

การทดลองเพื่อทำนายผลการเรียนทั้ง 4 ครั้งนั้นพบว่าแบบจำลอง Logistic Regression ให้ประสิทธิภาพการทำงานที่ดีที่สุดสำหรับการทำนายรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษ ในการทดลองการทำนายครั้งที่ 4, การทดลองการทำนายครั้งที่ 3 และการทดลองการทำนายครั้งที่ 4 ตามลำดับ ซึ่งสามารถแสดงรายละเอียดประสิทธิภาพได้ดังต่อไปนี้

ตาราง 19 แสดงรายละเอียดประสิทธิภาพการทำงานของแบบจำลอง Logistic Regression ในการทดลองการทำนายที่ 4 ในรายวิชาคณิตศาสตร์

คลาส	Precision	Recall	F1-Score
พอใช้	0.66	0.74	0.70
ดี	0.54	0.60	0.57
ดีมาก	0.89	0.78	0.83

จากตารางพบว่าแบบจำลอง Logistic Regression มีประสิทธิภาพการทำนายผลการเรียนรายวิชาคณิตศาสตร์ที่ดีที่สุดในคลาสดีมากที่ค่า precision 89% recall 78% และf1 83% รองลงมาคือคลาสพอใช้ที่ค่า precision 66% recall 74% และf1 70%และลำดับสุดท้ายคือคลาสดีที่ค่า precision 54% recall 60% และf1 57%ตามลำดับ



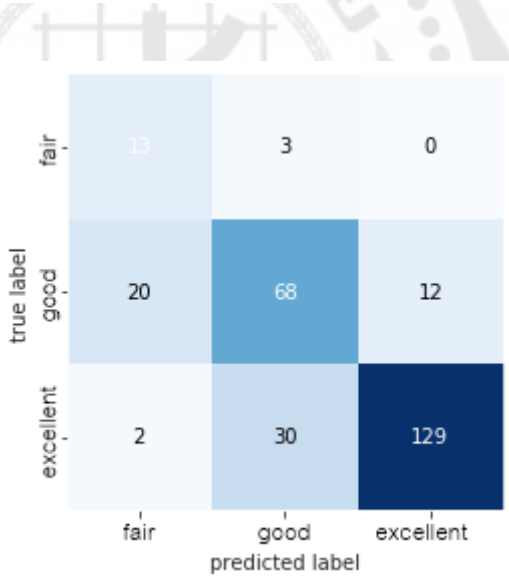
ภาพประกอบ 38 แสดงค่า confusion matrix ของแบบ Logistic Regression ในการทดลองการทำนายที่ 4 ในรายวิชาคณิตศาสตร์

จากค่า confusion matrix พบว่าประสิทธิภาพการทำนายของแบบจำลอง Logistic Regression ในรายวิชาคณิตศาสตร์สามารถทำนายได้ถูกต้องในคลาสพอใช้ 40 จาก 60 คน ทำนายได้ถูกต้องในคลาสดี 40 จาก 69 คน และทำนายได้ถูกต้องในคลาสดีมาก 60 จาก 87 คน

ตาราง 20 แสดงรายละเอียดประสิทธิภาพการทำงานของแบบจำลอง Logistic Regression ในการทดลองการทำนายที่ 3 ในรายวิชาวิทยาศาสตร์

คลาส	Precision	Recall	F1-Score
พอใช้	0.43	0.64	0.51
ดี	0.49	0.51	0.50
ดีมาก	0.82	0.71	0.76

จากตารางพบว่าแบบจำลอง Logistic Regression มีประสิทธิภาพการทำนายผลการเรียนรายวิชาวิทยาศาสตร์ที่ดีที่สุดในคลาสดีมากที่ค่า precision 82% recall 71% และf1 76% รองลงมาคือคลาสพอใช้ที่ค่า precision 43% recall 64% และf1 51% และลำดับสุดท้ายคือคลาสดีที่ค่า precision 49% recall 51% และf1 50% ตามลำดับ



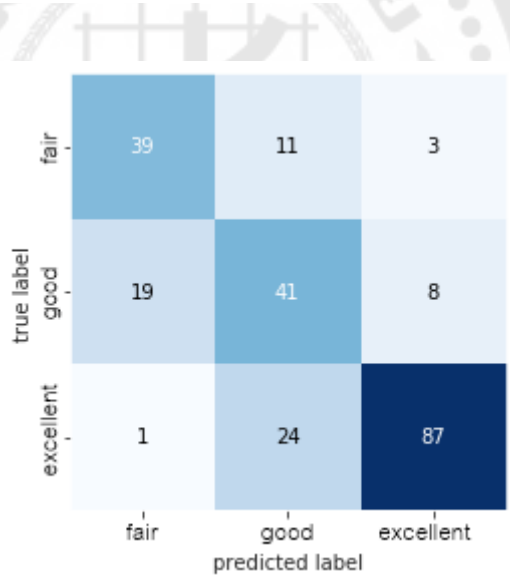
ภาพประกอบ 39 แสดงค่า confusion matrix ของแบบ Logistic Regression ในการทดลองการทำนายที่ 3 ในรายวิชาวิทยาศาสตร์

จากค่า confusion matrix พบว่าประสิทธิภาพการทำนายของแบบจำลอง Logistic Regression ในรายวิชาวิทยาศาสตร์สามารถทำนายได้ถูกต้องในคลาสพอใช้ 13 จาก 16 คน ทำนายได้ถูกต้องในคลาสดี 68 จาก 100 คน และทำนายได้ถูกต้องในคลาสดีมาก 129 จาก 161 คน

ตาราง 21 แสดงรายละเอียดประสิทธิภาพการทำงานของแบบจำลอง Logistic Regression ในการทดลองการทำนายที่ 4 ในรายวิชาภาษาอังกฤษ

คลาส	Precision	Recall	F1-Score
พอใช้	0.66	0.74	0.70
ดี	0.54	0.60	0.57
ดีมาก	0.89	0.78	0.83

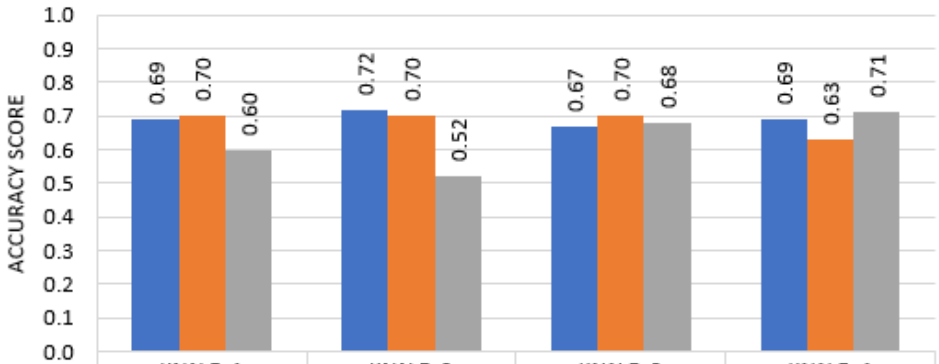
จากตารางพบว่าแบบจำลอง Logistic Regression มีประสิทธิภาพการทำนายผลการเรียนรายวิชาภาษาอังกฤษที่ดีที่สุดในคลาสดีมากที่ค่า precision 89% recall 78% และf1 83% รองลงมาคือคลาสพอใช้ที่ค่า precision 66% recall 74% และf1 70% และลำดับสุดท้ายคือคลาสดีที่ค่า precision 54% recall 60% และf1 57% ตามลำดับ



ภาพประกอบ 40 แสดงค่า confusion matrix ของแบบ Logistic Regression ในการทดลองการทำนายที่ 4 ในรายวิชาภาษาอังกฤษ

จากค่า confusion matrix พบว่าประสิทธิภาพการทำนายของแบบจำลอง Logistic Regression ในรายวิชาภาษาอังกฤษสามารถทำนายได้ถูกต้องในคลาสพอใช้ 39 จาก 53 คน ทำนายได้ถูกต้องในคลาสดี 41 จาก 68 คน และทำนายได้ถูกต้องในคลาสดีมาก 87 จาก 112 คน

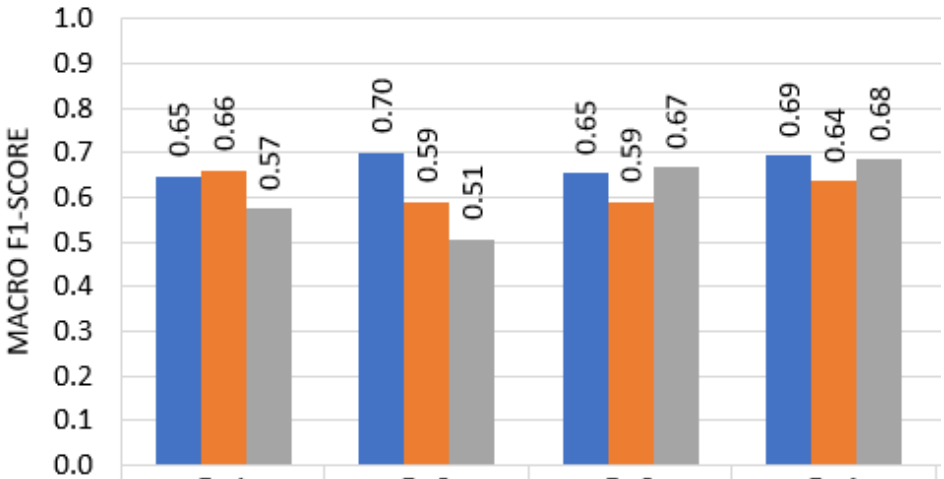
ผลลัพธ์ของการสร้างแบบจำลอง K-Nearest Neighbor



	KNN Ex1	KNN Ex2	KNN Ex3	KNN Ex4
math	0.69	0.72	0.67	0.69
sci	0.70	0.70	0.70	0.63
eng	0.60	0.52	0.68	0.71

ภาพประกอบ 41 แสดงผลการทำนายผลการเรียนของนักเรียน
ด้วยแบบจำลอง K-Nearest Neighbor

จากรูปภาพที่ 41 แสดงผลการทำนายผลการเรียนของนักเรียนด้วยแบบจำลอง K-Nearest Neighbor จากการทดลองทั้ง 4 ครั้ง พบว่าในการทดลองครั้งที่ 1 แบบจำลองได้ค่า accuracy ในรายวิชาคณิตศาสตร์ที่ 0.69 รายวิชาวิทยาศาสตร์ 0.70 และรายวิชาภาษาอังกฤษ 0.60 ต่อมาในการทดลองครั้งที่ 2 แบบจำลองได้ค่า accuracy ในรายวิชาคณิตศาสตร์ที่ 0.72 รายวิชาวิทยาศาสตร์ 0.70 และรายวิชาภาษาอังกฤษ 0.52 สำหรับการทดลองครั้งที่ 3 แบบจำลองได้ค่า accuracy ในรายวิชาคณิตศาสตร์ที่ 0.67 รายวิชาวิทยาศาสตร์ 0.70 และรายวิชาภาษาอังกฤษ 0.68 และในการทดลองครั้งที่ 4 แบบจำลองได้ค่า accuracy ในรายวิชาคณิตศาสตร์ที่ 0.69 รายวิชาวิทยาศาสตร์ 0.63 และรายวิชาภาษาอังกฤษ 0.71 ตามลำดับ



	Ex.1	Ex.2	Ex.3	Ex.4
math	0.65	0.70	0.65	0.69
sci	0.66	0.59	0.59	0.64
eng	0.57	0.51	0.67	0.68

ภาพประกอบ 42 แสดงค่า Macro F1-Score ในการทำนายผลการเรียนของนักเรียนด้วยแบบจำลอง K-Nearest Neighbor

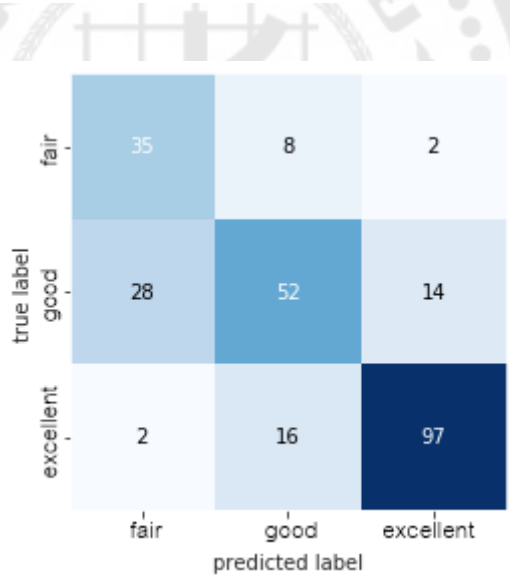
จากรูปภาพที่ 42 แสดงผล Macro F1-Score จากการทำนายผลการเรียนของนักเรียนด้วยแบบจำลอง K-Nearest Neighbor จากการทดลองทั้ง 4 ครั้ง พบว่าในการทดลองครั้งที่ 1 แบบจำลองได้ค่า Macro F1-Score รายวิชาคณิตศาสตร์ รายวิชาวิทยาศาสตร์ และรายวิชาภาษาอังกฤษ ที่ 0.65, 0.66 และ0.57 ตามลำดับ ในการทดลองครั้งที่ 2 ที่ค่า 0.70, 0.59 และ 0.51 ตามลำดับ ในการทดลองครั้งที่ 3 ที่ค่า 0.65, 0.59 และ0.67 ตามลำดับและในการทดลองครั้งที่ 4 ที่ค่า 0.69, 0.64 และ0.68 ตามลำดับ

การทดลองเพื่อทำนายผลการเรียนทั้ง 4 ครั้งนั้นพบว่าแบบจำลอง K-Nearest Neighbor ให้ประสิทธิภาพการทำงานที่ดีที่สุดสำหรับการทำนายรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษ ในการทดลองการทำนายครั้งที่ 2, การทดลองการทำนายครั้งที่ 1 และการทดลองการทำนายครั้งที่ 4 ตามลำดับ ซึ่งสามารถแสดงรายละเอียดประสิทธิภาพได้ดังต่อไปนี้

ตาราง 22 แสดงรายละเอียดประสิทธิภาพการทำงานของแบบจำลอง K-Nearest Neighbor ในการทดลองการทำนายที่ 2 ในรายวิชาคณิตศาสตร์

คลาส	Precision	Recall	F1-Score
พอใช้	0.54	0.78	0.64
ดี	0.68	0.55	0.61
ดีมาก	0.86	0.84	0.85

จากตารางพบว่าแบบจำลอง K-Nearest Neighbor มีประสิทธิภาพการทำนายผลการเรียนรายวิชาคณิตศาสตร์ที่ดีที่สุดในคลาสดีมากที่ค่า precision 86% recall 84% และf1 85% รองลงมาคือคลาสพอใช้ที่ค่า precision 54% recall 78% และf1 64%และลำดับสุดท้ายคือคลาสดีที่ค่า precision 68% recall 55% และf1 61%ตามลำดับ



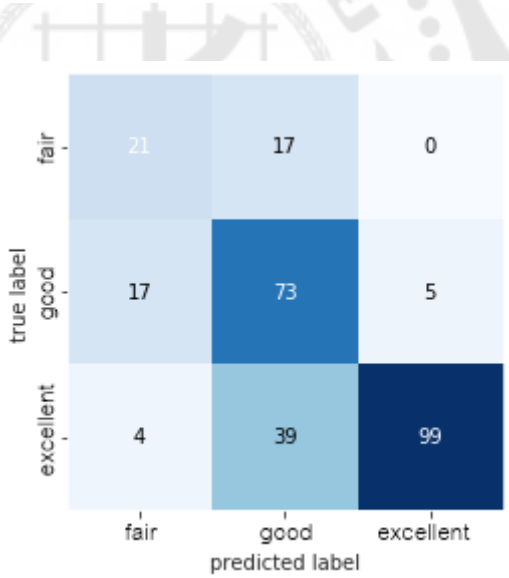
ภาพประกอบ 43 แสดงค่า confusion matrix ของแบบ K-Nearest Neighbor ในการทดลองการทำนายที่ 2 ในรายวิชาคณิตศาสตร์

จากค่า confusion matrix พบว่าประสิทธิภาพการทำนายของแบบจำลอง K-Nearest Neighbor ในรายวิชาคณิตศาสตร์สามารถทำนายได้ถูกต้องในคลาสพอใช้ 35 จาก 45 คน ทำนายได้ถูกต้องในคลาสดี 52 จาก 94 คน และทำนายได้ถูกต้องในคลาสดีมาก 97 จาก 115 คน

ตาราง 23 แสดงรายละเอียดประสิทธิภาพการทำงานของแบบจำลอง K-Nearest Neighbor ในการทดลองการทำนายที่ 1 ในรายวิชาวิทยาศาสตร์

คลาส	Precision	Recall	F1-Score
พอใช้	0.40	0.66	0.50
ดี	0.53	0.45	0.49
ดีมาก	0.82	0.76	0.79

จากตารางพบว่าแบบจำลอง K-Nearest Neighbor มีประสิทธิภาพการทำนายผลการเรียนรายวิชาวิทยาศาสตร์ที่ดีที่สุดในคลาสดีมากที่ค่า precision 82% recall 76% และf1 79% รองลงมาคือคลาสพอใช้ที่ค่า precision 40% recall 66% และf1 50% และลำดับสุดท้ายคือคลาสดีที่ค่า precision 53% recall 45% และf1 49% ตามลำดับ



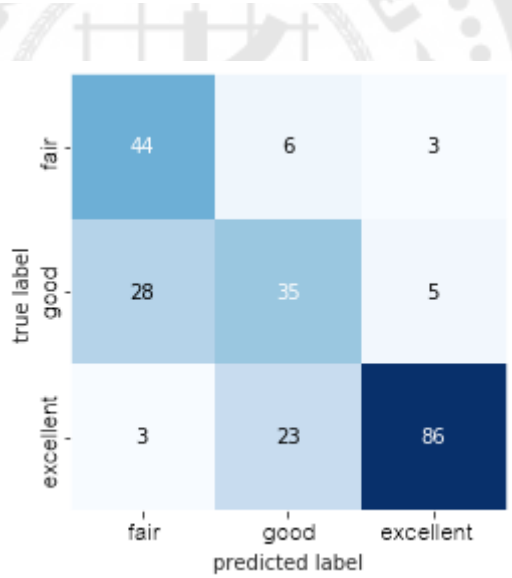
ภาพประกอบ 44 แสดงค่า confusion matrix ของแบบ K-Nearest Neighbor ในการทดลองการทำนายที่ 1 ในรายวิชาวิทยาศาสตร์

จากค่า confusion matrix พบว่าประสิทธิภาพการทำนายของแบบจำลอง K-Nearest Neighbor ในรายวิชาวิทยาศาสตร์สามารถทำนายได้ถูกต้องในคลาสพอใช้ 21 จาก 38 คน ทำนายได้ถูกต้องในคลาสดี 73 จาก 95 คน และทำนายได้ถูกต้องในคลาสดีมาก 99 จาก 142 คน

ตาราง 24 แสดงรายละเอียดประสิทธิภาพการทำงานของแบบจำลอง K-Nearest Neighbor ในการทดลองการทำนายที่ 4 ในรายวิชาภาษาอังกฤษ

คลาส	Precision	Recall	F1-Score
พอใช้	0.59	0.83	0.69
ดี	0.55	0.51	0.53
ดีมาก	0.91	0.77	0.83

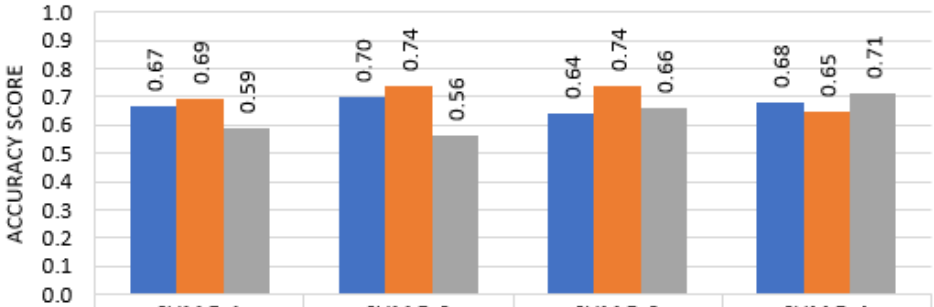
จากตารางพบว่าแบบจำลอง K-Nearest Neighbor มีประสิทธิภาพการทำนายผลการเรียนรายวิชาภาษาอังกฤษที่ดีที่สุดในคลาสดีมากที่ค่า precision 91% recall 77% และf1 83% รองลงมาคือคลาสพอใช้ที่ค่า precision 59% recall 83% และf1 69% และลำดับสุดท้ายคือคลาสดีที่ค่า precision 55% recall 51% และf1 53% ตามลำดับ



ภาพประกอบ 45 แสดงค่า confusion matrix ของแบบ K-Nearest Neighbor ในการทดลองการทำนายที่ 4 ในรายวิชาภาษาอังกฤษ

จากค่า confusion matrix พบว่าประสิทธิภาพการทำนายของแบบจำลอง K-Nearest Neighbor ในรายวิชาภาษาอังกฤษสามารถทำนายได้ถูกต้องในคลาสพอใช้ 44 จาก 53 คน ทำนายได้ถูกต้องในคลาสดี 35 จาก 68 คน และทำนายได้ถูกต้องในคลาสดีมาก 86 จาก 112 คน

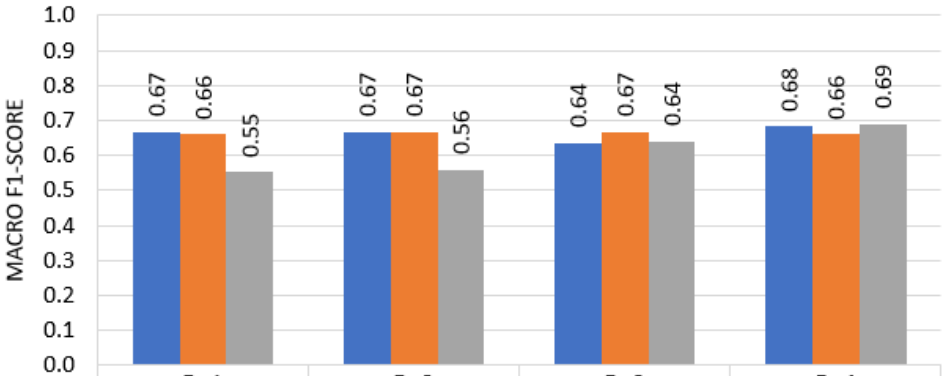
ผลลัพธ์ของการสร้างแบบจำลอง Support Vector Machines



	SVM Ex1	SVM Ex2	SVM Ex3	SVM Ex4
math	0.67	0.70	0.64	0.68
sci	0.69	0.74	0.74	0.65
eng	0.59	0.56	0.66	0.71

ภาพประกอบ 46 แสดงผลการทำนายผลการเรียนของนักเรียน
ด้วยแบบจำลอง Support Vector Machines

จากรูปภาพที่ 46 แสดงผลการทำนายผลการเรียนของนักเรียนด้วยแบบจำลอง Support Vector Machines จากการทดลองทั้ง 4 ครั้ง พบว่าในการทดลองครั้งที่ 1 แบบจำลองได้ค่า accuracy ในรายวิชาคณิตศาสตร์ที่ 0.67 รายวิชาวิทยาศาสตร์ 0.69 และรายวิชาภาษาอังกฤษ 0.59 ต่อมาในการทดลองครั้งที่ 2 แบบจำลองได้ค่า accuracy ในรายวิชาคณิตศาสตร์ที่ 0.70 รายวิชาวิทยาศาสตร์ 0.74 และรายวิชาภาษาอังกฤษ 0.56 สำหรับการทดลองครั้งที่ 3 แบบจำลองได้ค่า accuracy ในรายวิชาคณิตศาสตร์ที่ 0.64 รายวิชาวิทยาศาสตร์ 0.74 และรายวิชาภาษาอังกฤษ 0.66 และในการทดลองครั้งที่ 4 แบบจำลองได้ค่า accuracy ในรายวิชาคณิตศาสตร์ที่ 0.68 รายวิชาวิทยาศาสตร์ 0.65 และรายวิชาภาษาอังกฤษ 0.71 ตามลำดับ



	Ex.1	Ex.2	Ex.3	Ex.4
math	0.67	0.67	0.64	0.68
sci	0.66	0.67	0.67	0.66
eng	0.55	0.56	0.64	0.69

ภาพประกอบ 47 แสดงค่า Macro F1-Score ในการทำนายผลการเรียนของนักเรียนด้วยแบบจำลอง Support Vector Machines

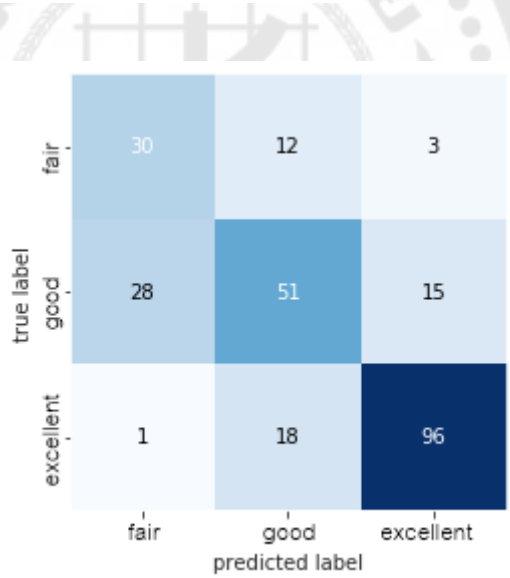
จากรูปภาพที่ 47 แสดงผล Macro F1-Score จากการทำนายผลการเรียนของนักเรียนด้วยแบบจำลอง Support Vector Machines จากการทดลองทั้ง 4 ครั้ง พบว่าในการทดลองครั้งที่ 1 แบบจำลองได้ค่า Macro F1-Score รายวิชาคณิตศาสตร์ รายวิชาวิทยาศาสตร์ และรายวิชาภาษาอังกฤษ ที่ 0.67, 0.66 และ0.55 ตามลำดับ ในการทดลองครั้งที่ 2 ที่ค่า 0.67, 0.67 และ 0.56 ตามลำดับ ในการทดลองครั้งที่ 3 ที่ค่า 0.64, 0.67 และ0.64 ตามลำดับและในการทดลองครั้งที่ 4 ที่ค่า 0.68, 0.66 และ0.69 ตามลำดับ

การทดลองเพื่อทำนายผลการเรียนทั้ง 4 ครั้งนั้นพบว่าแบบจำลอง Support Vector Machines ให้ประสิทธิภาพการทำงานที่ดีที่สุดสำหรับการทำนายรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษ ในการทดลองการทำนายครั้งที่ 2, การทดลองการทำนายครั้งที่ 2 และการทดลองการทำนายครั้งที่ 4 ตามลำดับ ซึ่งสามารถแสดงรายละเอียดประสิทธิภาพได้ดังต่อไปนี้

ตาราง 25 แสดงรายละเอียดประสิทธิภาพการทำงานของแบบจำลอง Support Vector Machines ในการทดลองการทำนายที่ 2 ในรายวิชาคณิตศาสตร์

คลาส	Precision	Recall	F1-Score
พอใช้	0.51	0.67	0.58
ดี	0.63	0.54	0.58
ดีมาก	0.84	0.83	0.84

จากตารางพบว่าแบบจำลอง Support Vector Machines มีประสิทธิภาพการทำนายผลการเรียนรายวิชาคณิตศาสตร์ที่ดีที่สุดในคลาสดีมากที่ค่า precision 84% recall 83% และf1 84% รองลงมาคือคลาสพอใช้ที่ค่า precision 51% recall 67% และf1 58%และลำดับสุดท้ายคือคลาสดีที่ค่า precision 63% recall 54% และf1 58%ตามลำดับ



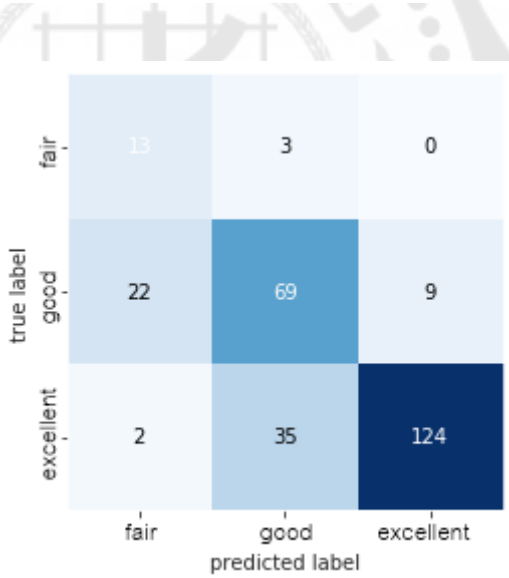
ภาพประกอบ 48 แสดงค่า confusion matrix ของแบบ Support Vector Machines ในการทดลองการทำนายที่ 2 ในรายวิชาคณิตศาสตร์

จากค่า confusion matrix พบว่าประสิทธิภาพการทำนายของแบบจำลอง Support Vector Machines ในรายวิชาคณิตศาสตร์สามารถทำนายได้ถูกต้องในคลาสพอใช้ 30 จาก 45 คน ทำนายได้ถูกต้องในคลาสดี 51 จาก 94 คน และทำนายได้ถูกต้องในคลาสดีมาก 96 จาก 115 คน

ตาราง 26 แสดงรายละเอียดประสิทธิภาพการทำงานของแบบจำลอง Support Vector Machines ในการทดลองการทำนายที่ 2 ในรายวิชาวิทยาศาสตร์

คลาส	Precision	Recall	F1-Score
พอใช้	0.35	0.81	0.49
ดี	0.64	0.69	0.67
ดีมาก	0.93	0.77	0.84

จากตารางพบว่าแบบจำลอง Support Vector Machines มีประสิทธิภาพการทำนายผลการเรียนรายวิชาวิทยาศาสตร์ที่ดีที่สุดในคลาสดีมากที่ค่า precision 93% recall 77% และf1 84% รองลงมาคือคลาสดีที่ค่า precision 64% recall 69% และf1 67% และลำดับสุดท้ายคือคลาสพอใช้ที่ค่า precision 35% recall 81% และf1 49% ตามลำดับ



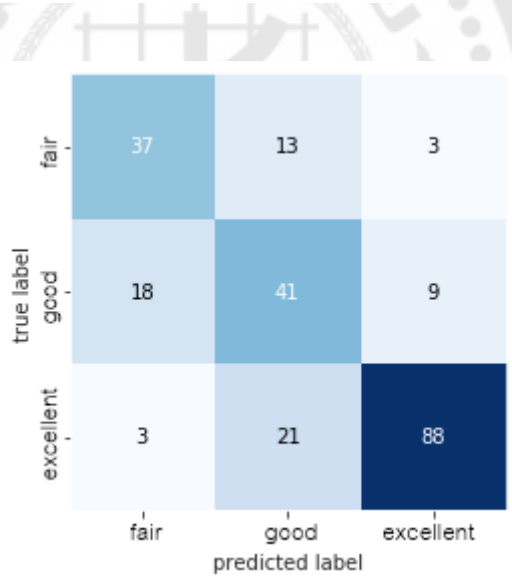
ภาพประกอบ 49 แสดงค่า confusion matrix ของแบบ Support Vector Machines ในการทดลองการทำนายที่ 2 ในรายวิชาวิทยาศาสตร์

จากค่า confusion matrix พบว่าประสิทธิภาพการทำนายของแบบจำลอง Support Vector Machines ในรายวิชาวิทยาศาสตร์สามารถทำนายได้ถูกต้องในคลาสพอใช้ 13 จาก 16 คน ทำนายได้ถูกต้องในคลาสดี 69 จาก 100 คน และทำนายได้ถูกต้องในคลาสดีมาก 124 จาก 161 คน

ตาราง 27 แสดงรายละเอียดประสิทธิภาพการทำงานของแบบจำลอง Support Vector Machines ในการทดลองการทำนายที่ 4 ในรายวิชาภาษาอังกฤษ

คลาส	Precision	Recall	F1-Score
พอใช้	0.64	0.70	0.67
ดี	0.55	0.60	0.57
ดีมาก	0.88	0.79	0.83

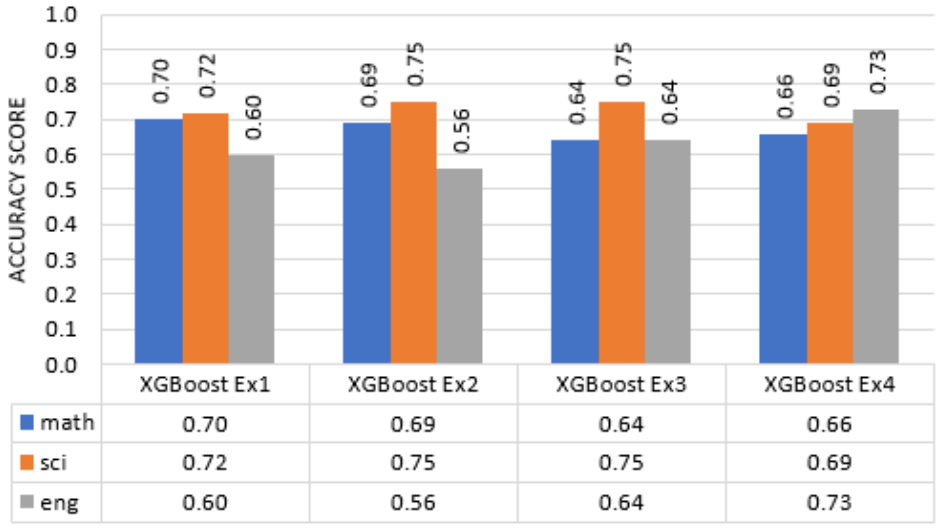
จากตารางพบว่าแบบจำลอง Support Vector Machines มีประสิทธิภาพการทำนายผลการเรียนรายวิชาภาษาอังกฤษที่ดีที่สุดในคลาสดีมากที่ค่า precision 88% recall 79% และf1 83% รองลงมาคือคลาสพอใช้ที่ค่า precision 64% recall 70% และf1 67% และลำดับสุดท้ายคือคลาสดีที่ค่า precision 55% recall 60% และf1 57% ตามลำดับ



ภาพประกอบ 50 แสดงค่า confusion matrix ของแบบ Support Vector Machines ในการทดลองการทำนายที่ 4 ในรายวิชาภาษาอังกฤษ

จากค่า confusion matrix พบว่าประสิทธิภาพการทำนายของแบบจำลอง Support Vector Machines ในรายวิชาภาษาอังกฤษสามารถทำนายได้ถูกต้องในคลาสพอใช้ 37 จาก 53 คน ทำนายได้ถูกต้องในคลาสดี 41 จาก 68 คน และทำนายได้ถูกต้องในคลาสดีมาก 88 จาก 112 คน

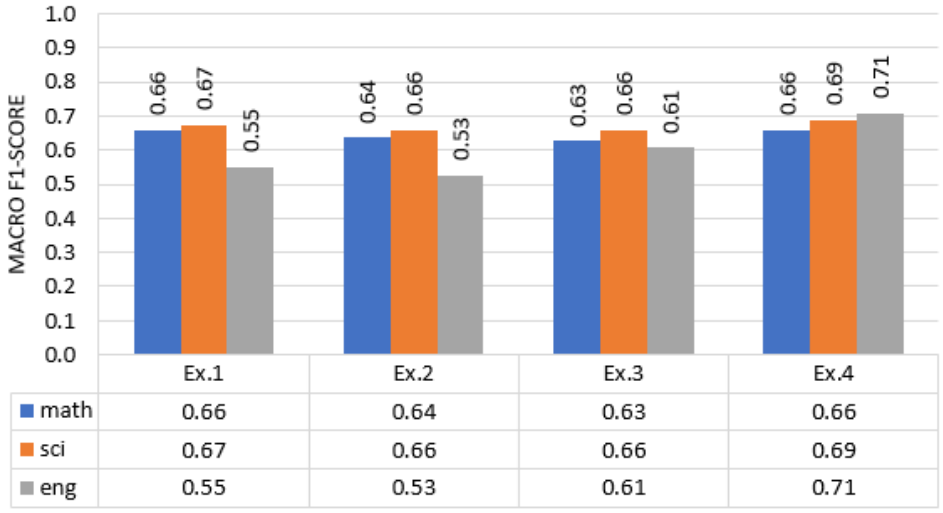
ผลลัพธ์ของการสร้างแบบจำลอง Extreme Gradient Boosting



	XGBoost Ex1	XGBoost Ex2	XGBoost Ex3	XGBoost Ex4
math	0.70	0.69	0.64	0.66
sci	0.72	0.75	0.75	0.69
eng	0.60	0.56	0.64	0.73

ภาพประกอบ 51 แสดงผลการทำนายผลการเรียนของนักเรียน
ด้วยแบบจำลอง Extreme Gradient Boosting

จากรูปภาพที่ 51 แสดงผลการทำนายผลการเรียนของนักเรียนด้วยแบบจำลอง Extreme Gradient Boosting จากการทดลองทั้ง 4 ครั้ง พบว่าในการทดลองครั้งที่ 1 แบบจำลองได้ค่า accuracy ในรายวิชาคณิตศาสตร์ที่ 0.70 รายวิชาวิทยาศาสตร์ 0.72 และรายวิชาภาษาอังกฤษ 0.60 ต่อมาในการทดลองครั้งที่ 2 แบบจำลองได้ค่า accuracy ในรายวิชาคณิตศาสตร์ที่ 0.69 รายวิชาวิทยาศาสตร์ 0.75 และรายวิชาภาษาอังกฤษ 0.56 สำหรับการทดลองครั้งที่ 3 แบบจำลองได้ค่า accuracy ในรายวิชาคณิตศาสตร์ที่ 0.64 รายวิชาวิทยาศาสตร์ 0.75 และรายวิชาภาษาอังกฤษ 0.64 และในการทดลองครั้งที่ 4 แบบจำลองได้ค่า accuracy ในรายวิชาคณิตศาสตร์ที่ 0.66 รายวิชาวิทยาศาสตร์ 0.69 และรายวิชาภาษาอังกฤษ 0.73 ตามลำดับ



	Ex.1	Ex.2	Ex.3	Ex.4
math	0.66	0.64	0.63	0.66
sci	0.67	0.66	0.66	0.69
eng	0.55	0.53	0.61	0.71

ภาพประกอบ 52 แสดงค่า Macro F1-Score ในการทำนายผลการเรียนของนักเรียน ด้วยแบบจำลอง Extreme Gradient Boosting

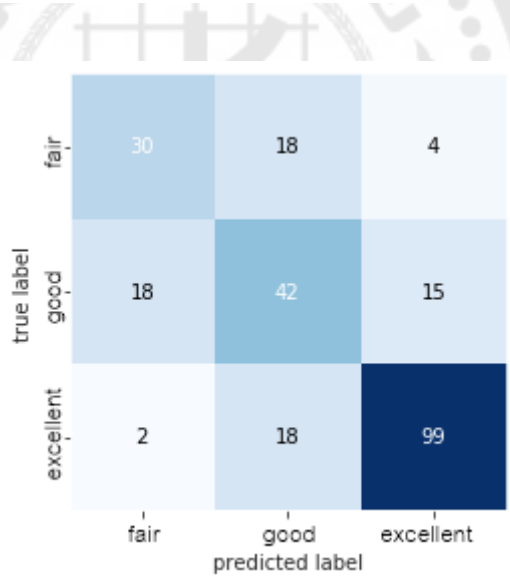
จากรูปภาพที่ 52 แสดงผล Macro F1-Score จากการทำนายผลการเรียนของนักเรียน ด้วยแบบจำลอง Extreme Gradient Boosting จากการทดลองทั้ง 4 ครั้ง พบว่าในการทดลองครั้งที่ 1 แบบจำลองได้ค่า Macro F1-Score รายวิชาคณิตศาสตร์ รายวิชาวิทยาศาสตร์ และรายวิชาภาษาอังกฤษ ที่ 0.66, 0.67 และ0.55 ตามลำดับ ในการทดลองครั้งที่ 2 ที่ค่า 0.64, 0.66 และ 0.53 ตามลำดับ ในการทดลองครั้งที่ 3 ที่ค่า 0.63, 0.66 และ0.61 ตามลำดับและในการทดลองครั้งที่ 4 ที่ค่า 0.66, 0.69 และ0.71 ตามลำดับ

การทดลองเพื่อทำนายผลการเรียนทั้ง 4 ครั้งนั้นพบว่าแบบจำลอง Extreme Gradient Boosting ให้ประสิทธิภาพการทำงานที่ดีที่สุดสำหรับการทำนายรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษ ในการทดลองการทำนายครั้งที่ 1, การทดลองการทำนายครั้งที่ 2 และการทดลองการทำนายครั้งที่ 4 ตามลำดับ ซึ่งสามารถแสดงรายละเอียดประสิทธิภาพได้ดังต่อไปนี้

ตาราง 28 แสดงรายละเอียดประสิทธิภาพการทำงานของแบบจำลอง Extreme Gradient Boosting ในการทดลองการทำนายที่ 1 ในรายวิชาคณิตศาสตร์

คลาส	Precision	Recall	F1-Score
พอใช้	0.60	0.58	0.59
ดี	0.54	0.56	0.55
ดีมาก	0.84	0.83	0.84

จากตารางพบว่าแบบจำลอง Extreme Gradient Boosting มีประสิทธิภาพการทำนายผลการเรียนรายวิชาคณิตศาสตร์ที่ดีที่สุดในคลาสดีมากที่ค่า precision 84% recall 83% และf1 84%รองลงมาคือคลาสพอใช้ที่ค่า precision 60% recall 58% และf1 59%และลำดับสุดท้ายคือคลาสดีที่ค่า precision 54% recall 56% และf1 55%ตามลำดับ



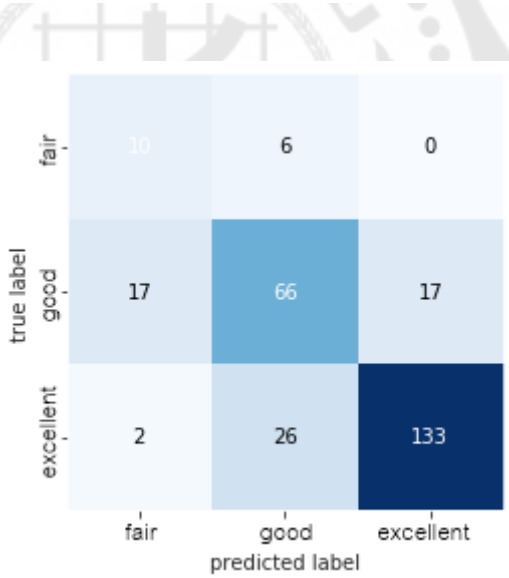
ภาพประกอบ 53 แสดงค่า confusion matrix ของแบบ Extreme Gradient Boosting ในการทดลองการทำนายที่ 1 ในรายวิชาคณิตศาสตร์

จากค่า confusion matrix พบว่าประสิทธิภาพการทำนายของแบบจำลอง Extreme Gradient Boosting ในรายวิชาคณิตศาสตร์สามารถทำนายได้ถูกต้องในคลาสพอใช้ 30 จาก 52 คน ทำนายได้ถูกต้องในคลาสดี 42 จาก 75 คน และทำนายได้ถูกต้องในคลาสดีมาก 99 จาก 119 คน

ตาราง 29 แสดงรายละเอียดประสิทธิภาพการทำงานของแบบจำลอง Extreme Gradient Boosting ในการทดลองการทำนายที่ 2 ในรายวิชาวิทยาศาสตร์

คลาส	Precision	Recall	F1-Score
พอใช้	0.34	0.62	0.44
ดี	0.67	0.66	0.67
ดีมาก	0.89	0.83	0.86

จากตารางพบว่าแบบจำลอง Extreme Gradient Boosting มีประสิทธิภาพการทำนายผลการเรียนรายวิชาวิทยาศาสตร์ที่ดีที่สุดในคลาสดีมากที่ค่า precision 89% recall 83% และf1 86% รองลงมาคือคลาสดีที่ค่า precision 67% recall 66% และf1 67% และลำดับสุดท้ายคือคลาสพอใช้ที่ค่า precision 34% recall 62% และf1 44% ตามลำดับ



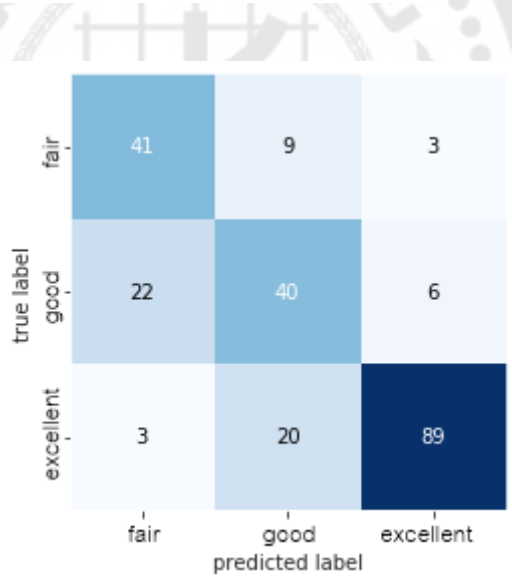
ภาพประกอบ 54 แสดงค่า confusion matrix ของแบบ Extreme Gradient Boosting ในการทดลองการทำนายที่ 2 ในรายวิชาวิทยาศาสตร์

จากค่า confusion matrix พบว่าประสิทธิภาพการทำนายของแบบจำลอง Extreme Gradient Boosting ในรายวิชาวิทยาศาสตร์สามารถทำนายได้ถูกต้องในคลาสพอใช้ 10 จาก 16 คน ทำนายได้ถูกต้องในคลาสดี 66 จาก 100 คน และทำนายได้ถูกต้องในคลาสดีมาก 133 จาก 161 คน

ตาราง 30 แสดงรายละเอียดประสิทธิภาพการทำงานของแบบจำลอง Extreme Gradient Boosting ในการทดลองการทำนายที่ 4 ในรายวิชาภาษาอังกฤษ

คลาส	Precision	Recall	F1-Score
พอใช้	0.62	0.77	0.69
ดี	0.58	0.59	0.58
ดีมาก	0.91	0.79	0.85

จากตารางพบว่าแบบจำลอง Extreme Gradient Boosting มีประสิทธิภาพการทำนายผลการเรียนรายวิชาภาษาอังกฤษที่ดีที่สุดในคลาสดีมากที่ค่า precision 91% recall 79% และf1 85% รองลงมาคือคลาสพอใช้ที่ค่า precision 62% recall 77% และf1 69% และลำดับสุดท้ายคือคลาสดีที่ค่า precision 58% recall 59% และf1 58% ตามลำดับ

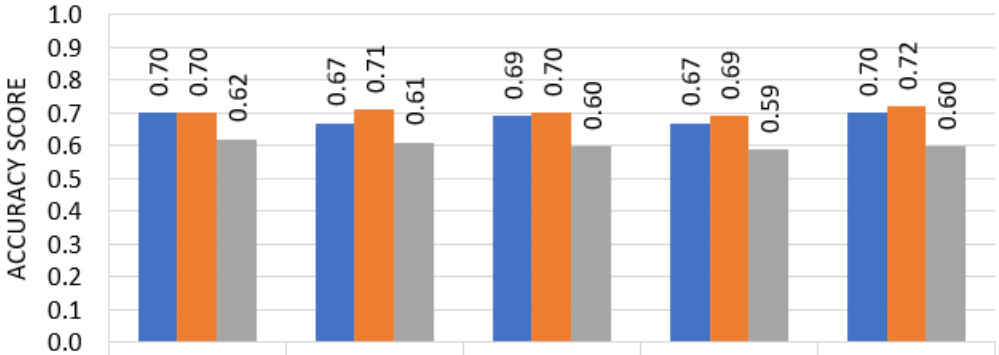


ภาพประกอบ 55 แสดงค่า confusion matrix ของแบบ Extreme Gradient Boosting ในการทดลองการทำนายที่ 4 ในรายวิชาภาษาอังกฤษ

จากค่า confusion matrix พบว่าประสิทธิภาพการทำนายของแบบจำลอง Extreme Gradient Boosting ในรายวิชาภาษาอังกฤษสามารถทำนายได้ถูกต้องในคลาสพอใช้ 41 จาก 53 คน ทำนายได้ถูกต้องในคลาสดี 40 จาก 68 คน และทำนายได้ถูกต้องในคลาสดีมาก 89 จาก 112 คน

ผลลัพธ์จากการเปรียบเทียบผลการทำนายผลการเรียนของนักเรียนของแบบจำลอง

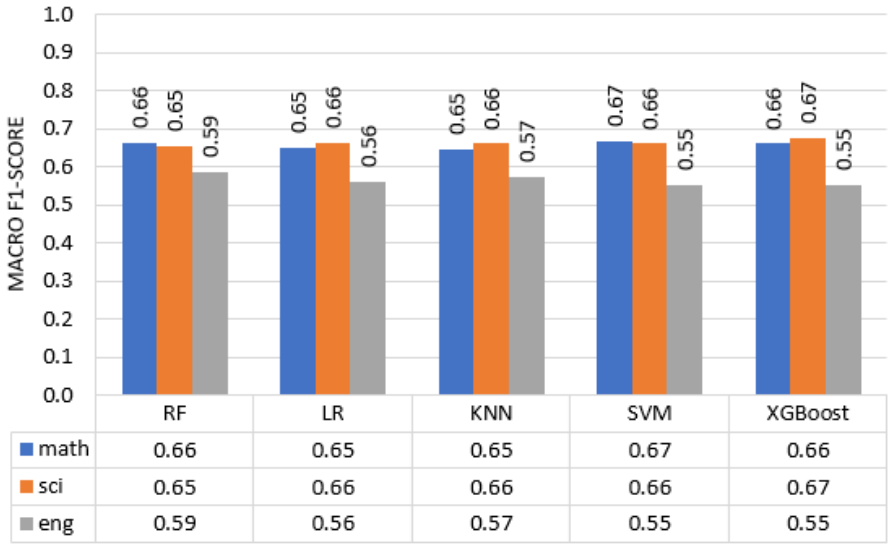
ผลลัพธ์ที่ได้จากการทดลองของแบบจำลอง Extreme Gradient Boosting, Logistic Regression, Support Vector Machine, K Nearest Neighbor และ Random Forest สามารถนำค่าที่ได้จากการทำนายมาเปรียบเทียบผลลัพธ์ในการทดลองครั้งที่ 1 ถึง ครั้งที่ 4 ดังนี้



	RF	LR	KNN	SVM	XGBoost
math	0.70	0.67	0.69	0.67	0.70
sci	0.70	0.71	0.70	0.69	0.72
eng	0.62	0.61	0.60	0.59	0.60

ภาพประกอบ 56 เปรียบเทียบผลการทำนายผลการเรียนของแบบจำลอง
ในการทดลองครั้งที่ 1 ด้วยค่า accuracy

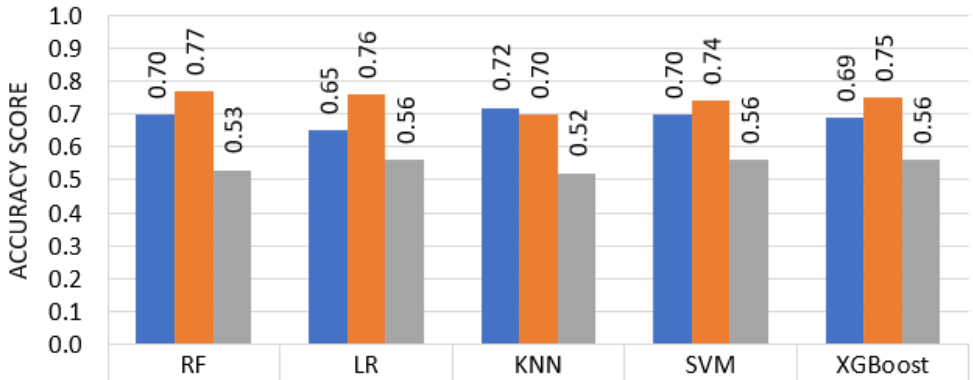
จากรูปภาพที่ 56 เปรียบเทียบค่า accuracy สำหรับการทำนายผลการเรียนของนักเรียนในการทดลองครั้ง 1 พบว่า รายวิชาคณิตศาสตร์แบบจำลองที่มีความแม่นยำมากที่สุด คือ Random Forest และ XGBoost ที่ค่า 0.70 และรายวิชาวิทยาศาสตร์แบบจำลองที่มีความแม่นยำมากที่สุด คือ XGBoost ที่ค่า 0.72 และรายวิชาภาษาอังกฤษแบบจำลองที่มีความแม่นยำมากที่สุด คือ Random Forest ที่ค่า 0.62



	RF	LR	KNN	SVM	XGBoost
math	0.66	0.65	0.65	0.67	0.66
sci	0.65	0.66	0.66	0.66	0.67
eng	0.59	0.56	0.57	0.55	0.55

ภาพประกอบ 57 เปรียบเทียบผลการทำนายผลการเรียนของแบบจำลอง
ในการทดลองครั้งที่ 1 ด้วยค่า Macro F1-Score

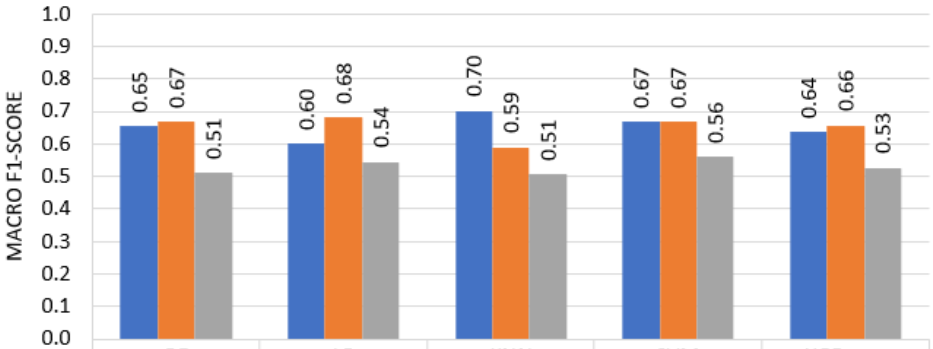
จากรูปภาพที่ 57 เปรียบเทียบค่า Macro F1-Score สำหรับการทำนายผลการเรียนของนักเรียนในการทดลองครั้ง 1 พบว่า รายวิชาคณิตศาสตร์แบบจำลองที่มีค่า Macro F1 มากที่สุด คือ Support Vector Machine ที่ค่า 0.67 ในรายวิชาวิทยาศาสตร์ได้แก่แบบจำลอง XGBoost ที่ค่า 0.67 และรายวิชาภาษาอังกฤษได้แก่แบบจำลอง Random Forest ที่ค่า 0.59



	RF	LR	KNN	SVM	XGBoost
math	0.70	0.65	0.72	0.70	0.69
sci	0.77	0.76	0.70	0.74	0.75
eng	0.53	0.56	0.52	0.56	0.56

ภาพประกอบ 58 เปรียบเทียบผลการทำนายผลการเรียนของแบบจำลอง
ในการทดลองครั้งที่ 2 ด้วยค่า accuracy

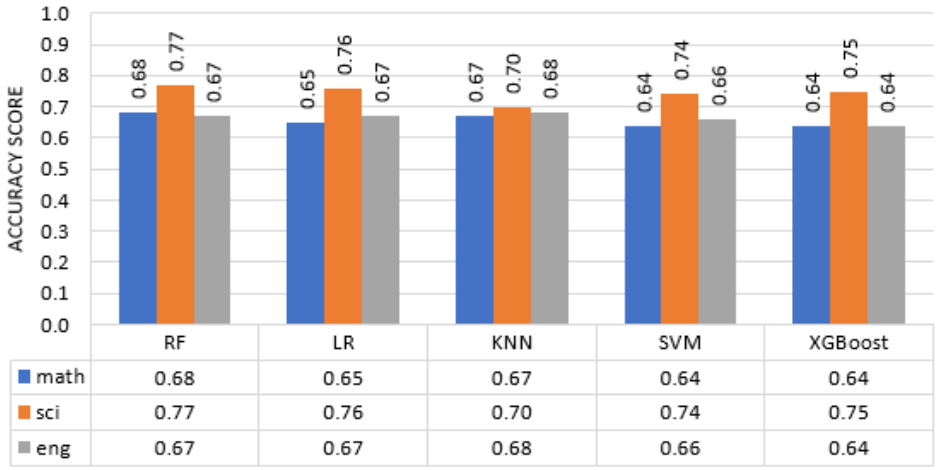
จากรูปภาพที่ 58 เปรียบเทียบค่า accuracy สำหรับการทำนายผลการเรียนของนักเรียนในการทดลองครั้ง 2 พบว่า รายวิชาคณิตศาสตร์แบบจำลองที่มีความแม่นยำมากที่สุด คือ K-Nearest Neighbor ที่ค่า 0.72 และรายวิชาวิทยาศาสตร์แบบจำลองที่มีความแม่นยำมากที่สุด คือ Random Forest ที่ค่า 0.77 และรายวิชาภาษาอังกฤษแบบจำลองที่มีความแม่นยำมากที่สุด คือ Support Vector Machines, Logistic Regression และ XGBoost ที่ค่า 0.56



	RF	LR	KNN	SVM	XGBoost
math	0.65	0.60	0.70	0.67	0.64
sci	0.67	0.68	0.59	0.67	0.66
eng	0.51	0.54	0.51	0.56	0.53

ภาพประกอบ 59 เปรียบเทียบผลการทำนายผลการเรียนของแบบจำลอง
ในการทดลองครั้งที่ *2* ด้วยค่า Macro F1-Score

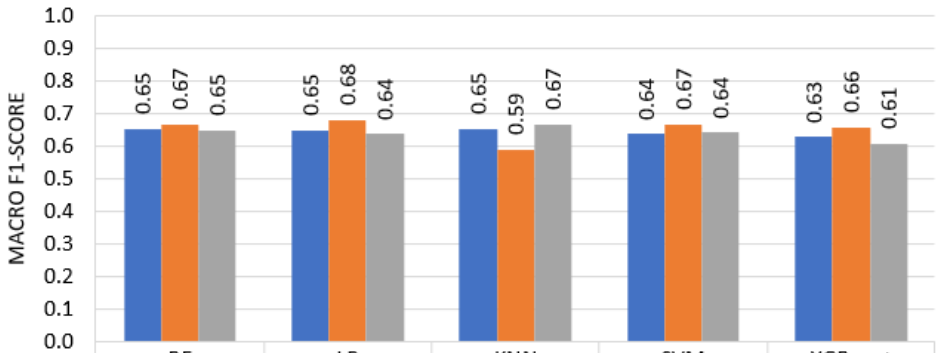
จากรูปภาพที่ 59 เปรียบเทียบค่า Macro F1-Score สำหรับการทำนายผลการเรียนของนักเรียนในการทดลองครั้ง 2 พบว่า รายวิชาคณิตศาสตร์แบบจำลองที่มีค่า Macro F1 มากที่สุด คือ K-Nearest Neighbor s ที่ค่า 0.70 ในรายวิชาวิทยาศาสตร์ได้แก่ แบบจำลอง Logistic Regression ที่ค่า 0.68 และรายวิชาภาษาอังกฤษคือ แบบจำลอง support vector machine ที่ค่า 0.56



	RF	LR	KNN	SVM	XGBoost
math	0.68	0.65	0.67	0.64	0.64
sci	0.77	0.76	0.70	0.74	0.75
eng	0.67	0.67	0.68	0.66	0.64

ภาพประกอบ 60 เปรียบเทียบผลการทำนายผลการเรียนของแบบจำลอง
ในการทดลองครั้งที่ 3 ด้วยค่า accuracy

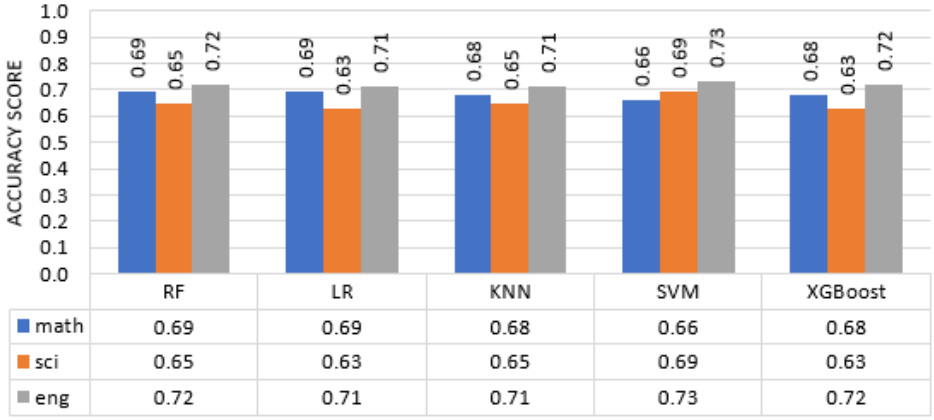
จากรูปภาพที่ 60 เปรียบเทียบค่า accuracy สำหรับการทำนายผลการเรียนของนักเรียนในการทดลองครั้ง 3 พบว่า รายวิชาคณิตศาสตร์แบบจำลองที่มีความแม่นยำมากที่สุด คือ Random Forest ที่ค่า 0.68 และรายวิชาวิทยาศาสตร์แบบจำลองที่มีความแม่นยำมากที่สุด คือ Random Forest ที่ค่า 0.77 และรายวิชาภาษาอังกฤษแบบจำลองที่มีความแม่นยำมากที่สุด คือ K-Nearest Neighbor ที่ค่า 0.68



	RF	LR	KNN	SVM	XGBoost
math	0.65	0.65	0.65	0.64	0.63
sci	0.67	0.68	0.59	0.67	0.66
eng	0.65	0.64	0.67	0.64	0.61

ภาพประกอบ 61 เปรียบเทียบผลการทำนายผลการเรียนของแบบจำลอง
ในการทดลองครั้งที่ 3 ด้วยค่า Macro F1-Score

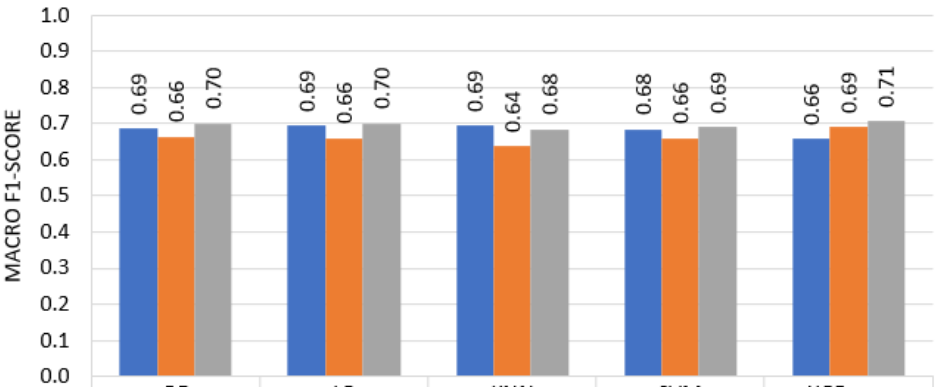
จากรูปภาพที่ 61 เปรียบเทียบค่า Macro F1-Score สำหรับการทำนายผลการเรียนของนักเรียนในการทดลองครั้ง 3 พบว่า รายวิชาคณิตศาสตร์แบบจำลองที่มีค่า Macro F1 มากที่สุด คือ Random Forest , Logistic Regression และ K-Nearest Neighbor s ที่ค่า 0.65 เท่ากัน ในรายวิชาวิทยาศาสตร์ได้แก่ แบบจำลอง Logistic Regression ที่ค่า 0.68 และรายวิชาภาษาอังกฤษ คือ แบบจำลอง K-Nearest Neighbors ที่ค่า 0.67



	RF	LR	KNN	SVM	XGBoost
math	0.69	0.69	0.68	0.66	0.68
sci	0.65	0.63	0.65	0.69	0.63
eng	0.72	0.71	0.71	0.73	0.72

ภาพประกอบ 62 เปรียบเทียบผลการทำนายผลการเรียนของแบบจำลอง
ในการทดลองครั้งที่ 4 ด้วยค่า accuracy

จากรูปภาพที่ 62 เปรียบเทียบค่า accuracy สำหรับการทำนายผลการเรียนของนักเรียนในการทดลองครั้ง 4 พบว่า รายวิชาคณิตศาสตร์แบบจำลองที่มีความแม่นยำมากที่สุด คือ Random Forest และ Logistic Regression ที่ค่า 0.69 และรายวิชาวิทยาศาสตร์แบบจำลองที่มีความแม่นยำมากที่สุด คือ Support Vector Machines ที่ค่า 0.69 และรายวิชาภาษาอังกฤษ แบบจำลองที่มีความแม่นยำมากที่สุด คือ Support Vector Machines ที่ค่า 0.73



	RF	LR	KNN	SVM	XGBoost
math	0.69	0.69	0.69	0.68	0.66
sci	0.66	0.66	0.64	0.66	0.69
eng	0.70	0.70	0.68	0.69	0.71

ภาพประกอบ 63 เปรียบเทียบผลการทำนายผลการเรียนของแบบจำลอง
ในการทดลองครั้งที่ 4 ด้วยค่า Macro F1-Score

จากรูปภาพที่ 63 เปรียบเทียบค่า Macro F1-Score สำหรับการทำนายผลการเรียนของนักเรียนในการทดลองครั้ง 4 พบว่า รายวิชาคณิตศาสตร์แบบจำลองที่มี Macro F1 มากที่สุด คือ Random Forest , Logistic Regression และ K-Nearest Neighbour ที่ค่า 0.69 เท่ากัน และรายวิชาวิทยาศาสตร์แบบจำลองที่มีความถูกต้องมากที่สุด คือ XGBoost ที่ค่า 0.69 และรายวิชาภาษาอังกฤษแบบจำลองที่มีความถูกต้องมากที่สุด คือ XGBoost ที่ค่า 0.71

บทที่ 5
สรุป อภิปรายผล และข้อเสนอแนะ

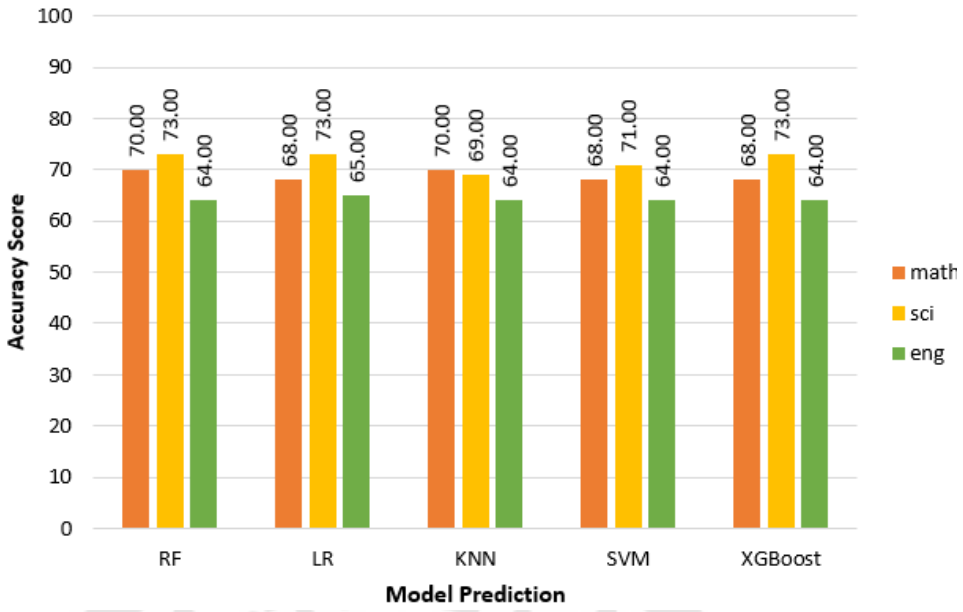
ในการวิจัยการทำนายผลการเรียนของนักเรียนในรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษ ซึ่งใช้ข้อมูลคะแนนรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษของนักเรียนจากโรงเรียนระดับมัธยมศึกษาแห่งหนึ่งในจังหวัดสุพรรณบุรี โดยใช้เทคนิคการเรียนรู้ของเครื่อง ผู้วิจัยได้วัดประสิทธิภาพของแบบจำลองแต่ละอัลกอริทึมเพื่อนำมาเปรียบเทียบและสรุปผล โดยสามารถแบ่งหัวข้อในการสรุปผลได้ดังนี้

1. สรุปผลการวิจัย
2. อภิปรายผลการวิจัย
3. ข้อเสนอแนะ

สรุปผลการวิจัย

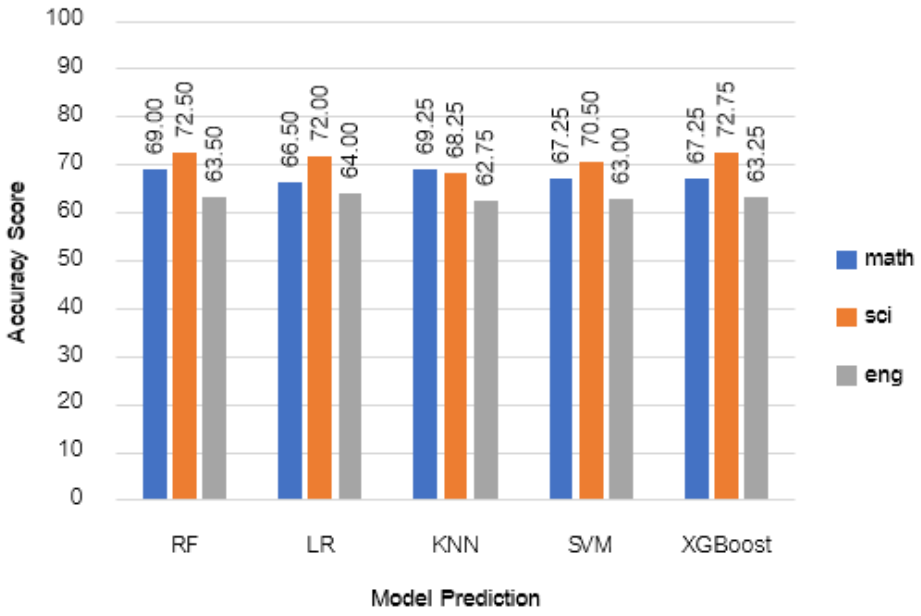
ปัจจุบันมีนักเรียนจำนวนมากประสบปัญหาการเลือกสาขาวิชาเรียนจากหลายสาเหตุ ซึ่งส่งผลกระทบทำให้นักเรียนไม่สามารถเลือกสาขาวิชาที่สนใจได้เนื่องจากมีผลการเรียนไม่เพียงต่อความต้องการของหลักสูตร ซึ่งสาขาวิชาวิศวกรรมศาสตร์หรือทางวิทยาศาสตร์และเทคโนโลยีนั้นถือเป็นสาขาวิชาที่ได้รับความสนใจอย่างมากสาขาหนึ่ง โดยจะพิจารณาผลการเรียนจากรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษเป็นสำคัญ ซึ่งปัญหาดังกล่าวนั้นสามารถที่จะประมาณการผลการเรียนของนักเรียนล่วงหน้าได้ว่านักเรียนจะมีผลการเรียนในแต่ละรายวิชาอยู่ในระดับใด ซึ่งจะช่วยให้นักเรียนสามารถปรับปรุงและวางแผนการเรียนของตนเองได้ นอกจากนี้ครูผู้สอนจะสามารถให้คำแนะนำและช่วยเหลือนักเรียนได้ทันเวลา

ในการวิจัยนี้ผู้วิจัยได้เลือกศึกษาข้อมูลของนักเรียนระดับมัธยมศึกษาตอนต้นจากโรงเรียนแห่งหนึ่งในจังหวัดสุพรรณบุรีและนำข้อมูลชุดดังกล่าวมาใช้ในการสร้างแบบจำลอง Extreme Gradient Boosting , Logistic Regression , Support Vector Machine, K Nearest Neighbor และ Random Forest โดยใช้เทคนิคการเรียนรู้ของเครื่อง เพื่อเป็นแนวทางในการพัฒนาการสร้างแบบจำลองในการทำนายผลการเรียนของนักเรียนในรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษ โดยสามารถสรุปการวัดประสิทธิภาพของแต่ละแบบจำลองได้ดังนี้



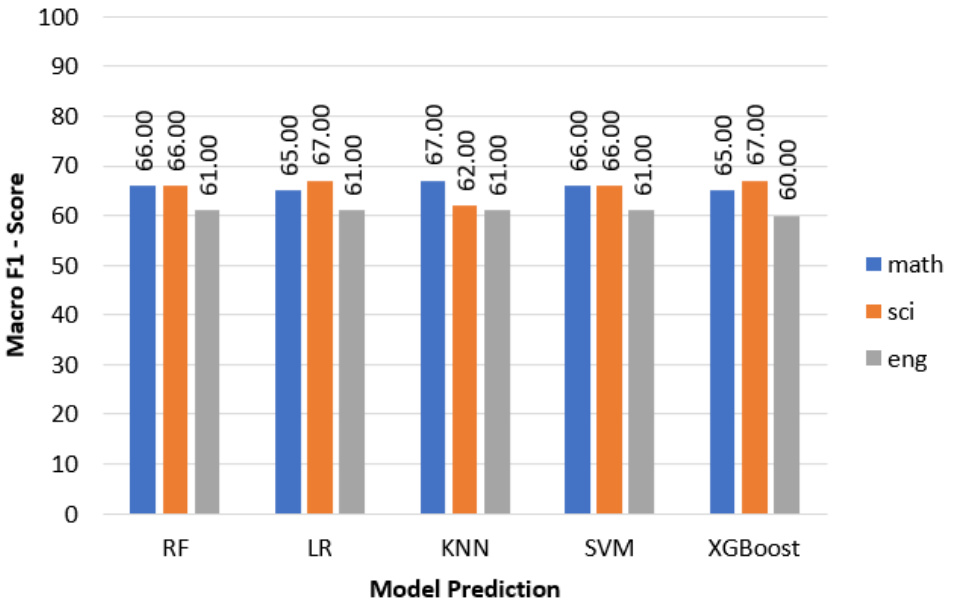
ภาพประกอบ 64 แสดงการเปรียบเทียบผลการทำนายเฉลี่ยสะสมด้วยค่า Accuracy ของแบบจำลองในการทำ 10 Folds Cross Validate

จากผลการทดลองสรุปได้ว่าแบบจำลองที่มีประสิทธิภาพเฉลี่ยสะสมดีที่สุดด้วยค่า accuracy สำหรับการ 10 Folds Cross Validate ในรายวิชาคณิตศาสตร์ คือ แบบจำลอง Random Forest และ K-Nearest Neighbor ที่ค่า 70% สำหรับการทำนายผลการเรียนรายวิชาวิทยาศาสตร์ คือ Random Forest, Logistic Regression และ XGBoost ที่ค่า 73% และอันดับที่ 3 คือ Support Vector Machines ที่ค่า 70.50% และสำหรับการทำนายผลการเรียนรายวิชาภาษาอังกฤษ คือ แบบจำลอง Logistic Regression ที่ค่า 65%



ภาพประกอบ 65 แสดงการเปรียบเทียบผลการทำนายเฉลี่ยสะสมด้วยค่า Accuracy ของแบบจำลองในการทำนายผลการเรียนของนักเรียน

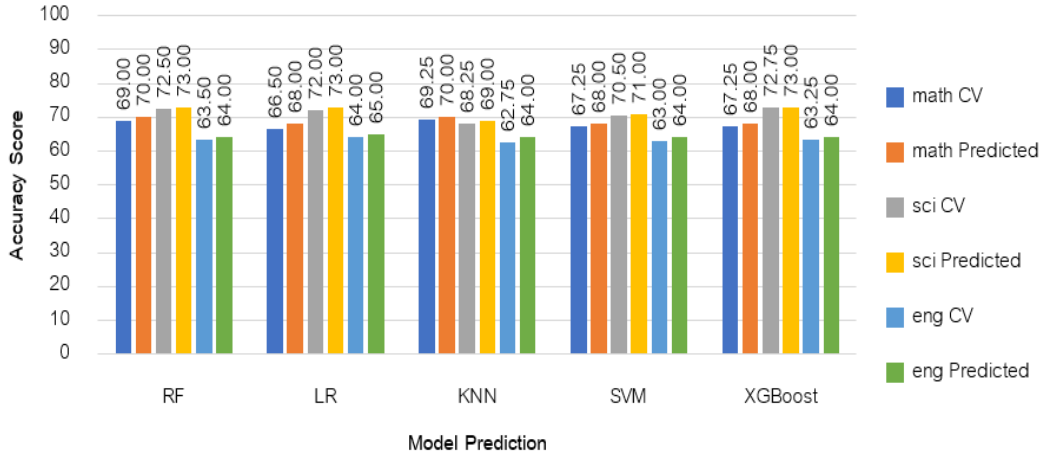
จากผลการทดลองสรุปได้ว่าแบบจำลองที่มีประสิทธิภาพเฉลี่ยสะสมดีที่สุดด้วยค่า accuracy สำหรับการทำนายผลการเรียนรายวิชาคณิตศาสตร์ คือ แบบจำลอง K-Nearest Neighbor ที่ค่า 69.25% อันดับที่ 2 คือ Random Forest ที่ค่า 69 % และอันดับที่ 3 คือ Support Vector Machines และ XGBoost ที่ค่า 67.25% สำหรับการทำนายผลการเรียนรายวิชาวิทยาศาสตร์ คือ แบบจำลอง XGBoost ที่ค่า 72.75% อันดับที่ 2 คือ Random Forest ที่ค่า 72.50 % และอันดับที่ 3 คือ Logistic Regression ที่ค่า 72% และสำหรับการทำนายผลการเรียนรายวิชาภาษาอังกฤษ คือ แบบจำลอง Logistic Regression ที่ค่า 64% อันดับที่ 2 คือ Random Forest ที่ค่า 63.50 % และอันดับที่ 3 คือ XGBoost ที่ค่า 63.25%



ภาพประกอบ 66 แสดงการเปรียบเทียบผลการทำนายเฉลี่ยสะสมด้วยค่า Macro F1-Score ของแบบจำลองในการทำนายผลการเรียนของนักเรียน

จากผลการทดลองสรุปได้ว่าแบบจำลองที่มีประสิทธิภาพเฉลี่ยสะสมดีที่สุดด้วยค่า Macro F1-Score สำหรับการทำนายผลการเรียนรายวิชาคณิตศาสตร์ คือ แบบจำลอง K-Nearest Neighbor ที่ค่า 67% อันดับที่ 2 คือ Random Forest และ Support Vector Machines ที่ค่า 66 % และอันดับที่ 3 คือ Logistic Regression และ XGBoost ที่ค่า 0.65% สำหรับการทำนายผลการเรียนรายวิชาวิทยาศาสตร์ คือ แบบจำลอง Logistic Regression และ XGBoost ที่ค่า 67% อันดับที่ 2 คือ Random Forest และ Support Vector Machines ที่ค่า 66 % และอันดับที่ 3 คือ K-Nearest Neighbor ที่ค่า 62% และสำหรับการทำนายผลการเรียนรายวิชาภาษาอังกฤษ แบบจำลอง K- Random Forest , Logistic Regression, K-Nearest Neighbor และ Support Vector Machines ได้ค่าเท่ากัน 61% และ XGBoost ที่ค่า 60%

จากประสิทธิภาพเฉลี่ยสะสมการทำงานของแบบจำลองเมื่อทำการเปรียบเทียบระหว่าง 10 Folds Cross Validate และการทำนายผลการเรียนของนักเรียนในรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษ พบว่า



ภาพประกอบ 67 แสดงประสิทธิภาพการทำงานเฉลี่ยสะสมระหว่าง 10 Folds Cross Validate และการทำนายผลการเรียนของนักเรียน

ประสิทธิภาพการทำงานเฉลี่ยสะสมในรายวิชาคณิตศาสตร์ผลลัพธ์จากการทำ 10 Folds Cross Validate ของแบบจำลอง Random Forest, Logistic Regression, K-Nearest Neighbor, Support Vector Machine และ Extreme Gradient Boosting ดังนี้ 69%, 66.50%, 69.25%, 67.25% และ 67.25% ตามลำดับ ในขณะที่ผลลัพธ์จากการทำนายผลการเรียนของนักเรียนอยู่ที่ค่า 70%, 68%, 70%, 68% และ 68% ตามลำดับ สำหรับรายวิชาวิทยาศาสตร์ได้ผลผลลัพธ์จากการทำ 10 Folds Cross Validate ที่ค่า 72.50%, 72%, 68.25%, 70.50% และ 72.75% ตามลำดับ ในขณะที่ผลลัพธ์จากการทำนายผลการเรียนของนักเรียนอยู่ที่ค่า 73%, 73%, 69%, 71% และ 73% ตามลำดับ และสำหรับรายวิชาภาษาอังกฤษได้ผลผลลัพธ์จากการทำ 10 Folds Cross Validate ที่ค่า 63.50%, 64%, 62.75%, 63% และ 63.25% ตามลำดับในขณะที่ผลลัพธ์จากการทำนายผลการเรียนของนักเรียนอยู่ที่ค่า 64%, 65%, 64%, 64% และ 64% ตามลำดับ

จากผลลัพธ์ข้างต้นแสดงถึงประสิทธิภาพการทำนายที่ใกล้เคียงกันระหว่างการทำ 10 Folds Cross Validate และการทำนายผลการเรียนของนักเรียน ซึ่งหมายถึงการทำงานของแบบจำลองที่ไม่เกิดปัญหา Overfiting Data

จากการทำนายผลการเรียนของนักเรียนเพื่อพิจารณาถึงสาเหตุของการทำนายคลาสระดับผลการเรียนที่ทำนายถูกและทำนายผิด ผู้วิจัยจึงวิเคราะห์ตัวอย่างข้อมูลที่ใช้ในการทำนายผลการเรียน ดังแสดงในตารางต่อไปนี้

ตาราง 31 แสดงตัวอย่างชุดข้อมูลการทดสอบที่ใช้ในการทำนายผลการเรียนของนักเรียน

Correct Prediction	Incorrect Prediction	Predicted Label	True Label
61, 50, 56, 68, 66, 67, 66, 70, 68	50, 61, 56 74, 80, 77, 65, 67, 66	ดี	พอใช้
55, 48, 52, 62, 70, 66, 61, 60, 61	76, 60, 68, 71, 79, 75, 78, 76, 77	ดี	พอใช้
72, 51, 62, 73, 68, 71, 62, 71, 67	72, 69, 71, 68, 66, 67, 76, 88, 82	ดี	พอใช้
51, 51, 51, 50, 60, 55, 61, 56, 59	77, 69, 73, 73, 82, 78, 70, 70, 70	ดีมาก	พอใช้
63, 60, 62, 57, 66, 62, 70, 65, 68	70, 77, 74, 80, 80, 80, 87, 87, 87	ดีมาก	พอใช้
61, 67, 64, 64, 61, 63, 71, 58, 65	84, 81, 83, 81, 87, 84, 80, 77, 79	ดีมาก	พอใช้
61, 82, 72 72, 72, 72, 70, 68, 69	69, 47, 58, 62, 71, 67, 81, 86, 84	พอใช้	ดี
66, 76, 71, 68, 73, 71 75, 62, 69	68, 52, 60, 72, 71, 72, 58, 62, 60	พอใช้	ดี
60, 70, 65, 76, 72, 74, 73, 77, 75	56, 51, 54, 78, 71, 75, 67, 60, 64	พอใช้	ดี
68, 66, 67, 67, 69, 68, 74, 65, 70	82, 76, 79, 75, 89, 82, 85, 84, 85	ดีมาก	ดี
65, 65, 65, 72, 75, 74, 80, 72, 76	75, 88, 82, 81, 80, 81, 71, 71, 71	ดีมาก	ดี
73, 61, 67, 71, 76, 74, 82, 76, 79	77, 76, 77, 80, 88, 84, 89, 92, 91	ดีมาก	ดี
83, 82, 83, 86, 90, 88, 93, 93, 93	71, 84, 78, 76, 71, 74, 76, 76, 76	ดี	ดีมาก
70, 72, 71, 81, 76, 79, 78, 79, 79	65, 52, 59 81, 77, 79, 74, 71, 73	ดี	ดีมาก
80, 87, 84, 85, 88, 87, 89, 87, 88	75, 72, 74, 75, 71, 73, 80, 72, 76,	ดี	ดีมาก
70, 72, 71, 77, 82, 80, 83, 75, 79	70, 50, 60, 66, 73, 70, 72, 75, 74	พอใช้	ดีมาก
86, 72, 79, 80, 75, 78, 85, 76, 81	50, 51, 51, 50, 69, 60, 65, 52, 59	พอใช้	ดีมาก
73, 73, 73, 75, 81, 78, 85, 91, 88	66, 51, 59, 78, 73, 76, 73, 60, 67	พอใช้	ดีมาก

จากการทำนายผลการเรียนพบว่าตัวอย่างข้อมูลที่แบบจำลองทำนายคลาสได้ถูกต้องนั้นมีช่วงคะแนนเป็นไปตามเกณฑ์ของระดับผลการเรียนที่กำหนด แต่ตัวอย่างข้อมูลที่แบบจำลองทำนายคลาสผิดนั้นช่วงคะแนนบางส่วนมีความใกล้เคียงกันมากในแต่ละคลาสทำให้แบบจำลอง

อาจแยกคลาสได้ยาก อีกทั้งพบว่าข้อมูลคะแนนรายวิชาเพียงอย่างเดียวนั้นอาจไม่เพียงพอต่อการทำนายผลการเรียน เนื่องจากข้อมูลตัวอย่างบางส่วนนั้นมีค่าคะแนนที่แบบจำลองควรทำนายได้ถูกต้องหากเทียบกับเกณฑ์ที่กำหนด แต่ผลการเรียนจริงพบว่านักเรียนได้ผลการเรียนที่ต่างออกไป ดังนั้นสามารถสรุปได้ว่าคะแนนในอดีตอาจไม่ใช่คุณลักษณะที่สำคัญเพียงอย่างเดียวที่จะใช้ทำนาย เพราะการที่นักเรียนมีผลการเรียนในอดีตดีหรือแย่นั้นไม่สามารถนำมาเป็นตัวแปรที่ชี้ขาดผลการเรียนในปัจจุบันได้ จึงจำเป็นต้องเก็บข้อมูลอื่นๆเพิ่มเติมเพื่อประกอบการทำนายให้มีประสิทธิภาพมากยิ่งขึ้น

อภิปรายผลการวิจัย

จากการทำการทดลองการทำนายผลการเรียนรายวิชาคณิตศาสตร์ วิทยาศาสตร์ และภาษาอังกฤษของนักเรียนระดับชั้นมัธยมศึกษาตอนต้นทั้งหมด 4 ครั้งโดยใช้แบบจำลอง Extreme Gradient Boosting , Logistic Regression , Support Vector Machine, K Nearest Neighbor และ Random Forest นั้น พบว่าแบบจำลองทั้ง 5 แบบจำลอง สามารถเรียนรู้ข้อมูลที่มีลักษณะแบบ multiclass ได้ ดังนั้นจึงเหมาะกับการนำมาใช้ในการทำนายผลการเรียนของนักเรียนทั้ง 3 รายวิชา ซึ่งพบว่าแบบจำลองมีประสิทธิภาพการทำนายที่ใกล้เคียงกัน โดย feature ที่มีความสำคัญกับการทำนายผลการเรียนสำหรับกลุ่มรายวิชาคณิตศาสตร์ได้แก่ sci2, ave_math, math2, math3, mathave_2, math4_eve, math3_ave, math5 และ ave3_math ในส่วน feature ที่มีความสำคัญกับการทำนายผลการเรียนสำหรับกลุ่มรายวิชาวิทยาศาสตร์ ได้แก่ ave_sci, sci2, mathave_3, sci3, sciave_3, ave5_sci, ave4_sci และ math5 และ feature ที่มีความสำคัญกับการทำนายผลการเรียนสำหรับกลุ่มรายวิชาภาษาอังกฤษ ได้แก่ ave_sci, ave_eng, math2, eng3, engave_3, sci1, eng4_ave, eng4, math4, ave5_eng, ave2_eng และeng5

นอกจากนี้ยังพบว่าแบบจำลองทั้ง 5 แบบจำลองมีความสามารถในการทำนายผลการเรียนรายวิชาวิทยาศาสตร์ได้ดีที่สุด รองลงมาคือการทำนายผลการเรียนรายวิชาคณิตศาสตร์ และการทำนายผลการเรียนรายวิชาภาษาอังกฤษที่ได้ประสิทธิภาพน้อยที่สุด และแบบจำลองสามารถทำนายผลการเรียนในคลาสดีมากได้อย่างมีประสิทธิภาพมากที่สุด ในส่วนของคลาสดีและคลาสพอใช้นั้นได้ผลการทำนายที่ใกล้เคียงกัน ซึ่งการที่ความแม่นยำในการทำนาย รวมถึงค่า precision และ recall ในแต่ละคลาสมีความแตกต่างกันนั้น ทั้งนี้อาจเนื่องมาจากข้อมูลที่ใช้ทำนายมีเพียงข้อมูลคะแนนรายวิชาเท่านั้น ซึ่งอาจไม่เพียงพอต่อการให้การทำนายผลการเรียน จึงจำเป็นที่จะต้องเพิ่มคุณลักษณะอื่นๆประกอบการทำนาย เช่น ข้อมูลพฤติกรรมการเรียน ทักษะและความถนัดในการเรียน ความรับผิดชอบของนักเรียน เป็นต้น นอกจากนี้สาเหตุอื่นๆอาจมาจากความ

แตกต่างทางด้านลักษณะรายวิชา ซึ่งเป็นไปได้ว่านักเรียนอาจมีผลการเรียนในรายวิชาวิทยาศาสตร์ที่คล้ายกัน คือนักเรียนได้คะแนนเต็มในรายวิชาดังกล่าว หรือผลการเรียนของนักเรียนอยู่ในเกณฑ์ดีมากทำให้แบบจำลองสามารถทำนายผลการเรียนได้ง่ายกว่าในรายวิชาคณิตศาสตร์และภาษาอังกฤษที่มีลักษณะของข้อมูลค่อนข้างแตกต่างและกระจายกันไปในแต่ละบุคคล ส่งผลให้แบบจำลองทำนายผลการเรียนได้ความแม่นยำน้อยกว่า อีกทั้งยังคงมีความแตกต่างทางด้านพื้นฐานของนักเรียนกล่าวคือ นักเรียนที่มีความสามารถและผลสัมฤทธิ์ใกล้เคียงกันในคลาสดีมากเป็นส่วนใหญ่ ย่อมส่งผลให้เกิดคุณลักษณะของนักเรียนในคลาสนั้นๆและทำให้แบบจำลองสามารถทำนายคลาสดีมากได้แม่นยำมากกว่าคลาสดี และคลาสพอใช้ที่มีข้อมูลแบบกระจายกันไปมากกว่า รวมถึงอาจารย์ผู้สอนที่แม้จะสอนในรายวิชาเดียวกันแต่อาจไม่ใช่คนเดียวกัน ซึ่งอาจต้องเพิ่ม feature การทำนายสำหรับปัจจัยด้านอื่นๆเพิ่มเติม รวมถึงเพิ่มจำนวนข้อมูลให้มากยิ่งขึ้นเพื่อพัฒนาแบบจำลองให้มีประสิทธิภาพสูงขึ้น เพื่อช่วยในการทำนายผลการเรียนของนักเรียนให้มีความแม่นยำและมีประสิทธิภาพเพิ่มมากขึ้น

ข้อเสนอแนะ

1. เนื่องจากในการวิจัยนี้มีจำนวนข้อมูลค่อนข้างน้อย ทำให้ข้อมูลสำหรับใช้ในการเรียนรู้อาจไม่เพียงพอ ดังนั้นหากสามารถเพิ่มจำนวนข้อมูล อาจช่วยเพิ่มประสิทธิภาพในการทำนายผลการเรียนของแบบจำลองให้ดียิ่งขึ้น

2. ในการวิจัยนี้ได้ใช้ข้อมูลในการศึกษาจากสถานศึกษาเพียงแห่งเดียว หากใช้ข้อมูลที่ถูกเก็บจากสถานศึกษาอื่นด้วย อาจช่วยให้แบบจำลองนั้นสามารถเรียนรู้ข้อมูลที่หลากหลายมากขึ้น รวมถึงสามารถทำนายค่าได้อย่างแม่นยำและมีประสิทธิภาพมากกว่าแบบจำลองที่ใช้ข้อมูลจากสถานศึกษาแห่งเดียว

3. เนื่องจากการวิจัยใช้ข้อมูลคะแนนรายวิชาเพียงอย่างเดียวในการทำนาย อาจทำให้การทำนายได้ผลที่แม่นยำไม่มากเท่าที่ควร หากมีการเพิ่มคุณลักษณะอื่นๆเพิ่มเติมอาจทำให้ผลการทำนายดีมากขึ้น เช่น ข้อมูลการเข้าเรียนของนักเรียน, ข้อมูลการภาระงาน, ข้อมูลจากการตอบแบบสอบถามด้านความถนัดเฉพาะด้าน หรือข้อมูลรายละเอียดของครูผู้สอน เป็นต้น

4. เนื่องจากในการวิจัยนี้ได้ศึกษาและใช้อัลกอริทึม 5 แบบ คือ Extreme Gradient Boosting, Logistic Regression , Support Vector Machine, K Nearest Neighbor และ Random Forest ดังนั้นอาจมีอัลกอริทึมอื่นที่สามารถเรียนรู้และทำนายค่าออกมาได้มีความแม่นยำและมีประสิทธิภาพดีกว่า

บรรณานุกรม

arnondora. (2019). feature selection machine learning. https://arnondora.in.th/feature-selection-machine-learning

Brink, H., Richards, J. W., และ Fetherolf, M. (2016). Capital Letter Real-World Machine Learning.

Chalermkiatsakul, P. (2018). Supervised Learning คืออะไร? ทำงานยังไง? https://medium.com/@every.phu/supervised-learning-%E0%B8%84%E0%B8%B7%E0%B8%AD%E0%B8%AD%E0%B8%B0%E0%B9%84%E0%B8%A3-%E0%B8%97%E0%B8%B3%E0%B8%87%E0%B8%B2%E0%B8%99%E0%B8%A2%E0%B8%B1%E0%B8%87%E0%B9%84%E0%B8%87-1c0e411a40a2

Chandrayan, P. (2015). Logistic Regression For Dummies: A Detailed Explanation. https://towardsdatascience.com/logistic-regression-for-dummies-a-detailed-explanation-9597f76edf46

Chen, X.-w., และ Jeong, J. C. (2008). *Enhanced recursive feature elimination*.

developers, s.-l. (1999). sklearn.feature_selection.RFE. https://ogrisel.github.io/scikit-learn.org/sklearn-tutorial/modules/generated/sklearn.feature_selection.RFE.html

Dorpe, S. V. (2018). Preprocessing with sklearn: a complete and comprehensive guide. https://towardsdatascience.com/preprocessing-with-sklearn-a-complete-and-comprehensive-guide-670cb98fcfb9

Iqbal, Z., Qadir, J., Mian, A., และ Kamiran, F. (2017). Machine Learning Based Student Grade Prediction: A Case Study.

Jordan, J. (2018). Learning from imbalanced data. https://www.jeremyjordan.me/imbalanced-data/

Latysheva, N. (2016). Implementing Your Own k-Nearest Neighbor Algorithm Using Python. https://www.kdnuggets.com/2016/01/implementing-your-own-knn-using-python.html

Meier, Y., Xu, J., Atan, O., และ Schaar, M. v. d. (2016). Predicting Grades. *IEEE Transactions on Signal Processing,* 64(4), 959-972.

Navlani, A. (2018). Support Vector Machines with Scikit-learn. https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python#svm

Odegua, R. (2015). FEATURE ENGINEERING AND DATA PREPARATION USING SUPERMARKET SALES DATA. https://towardsdatascience.com/feature-engineering-and-data-preparation-using-supermarket-sales-data-part-2-171b7a7a7eb7

Pathak, M. (2019). Using XGBoost in Python. https://www.datacamp.com/community/tutorials/xgboost-in-python

Polyzou, A., และ Karypis, G. (2018). *Feature extraction for classifying students based on their academic performance.* Paper presented at the Proceedings of the 11th International Conference on Educational Data Mining, EDM 2018.

Sweeney, M., Rangwala, H., Lester, J., และ Johri, A. (2016). Next-Term Student Performance Prediction: A Recommender Systems Approach.

Venkat, N., Srivastava, S., และ Garg, L. (2018). *Predicting Student Grades using Machine Learning*.

Vijite, P. (2018). ประเภทของ Machine Learning. https://medium.com/coeffest/%E0%B8%9B%E0%B8%A3%E0%B8%B0%E0%B9%80%E0%B8%A0%E0%B8%97%E0%B8%82%E0%B8%AD%E0%B8%87-machine-learning-f3159fee7b56

Wikipedia. (2019). Feature engineering. https://en.wikipedia.org/wiki/Feature_engineering

Zollanvari, A., Kizilirmak, R. C., Kho, Y. H., และ Hernández-Torrano, D. (2017). Predicting Students' GPA and Developing Intervention Strategies Based on Self-Regulatory Learning Behaviors. *IEEE Access,* 5, 23792-23802.

บุตรเอก, พ. (2014). การพยากรณ์โอกาสสำเร็จการศึกษาของนักศึกษา โดยใช้ซัพพอร์ตเวกเตอร์แมชชีน. *Veridian E-journal Science and Technology Silpakorn University,* 1(6), 40-49.

เลาหะวรนันท์, จ., ลิ้มสุทธิวันภูมิ, ร., ฐานะโสภณ, บ., และ เนติโสภากุล, พ. (2018). การใช้เทคนิคการทำเหมืองข้อมูลในการจำแนกและคัดเลือกแขนงวิชาสำหรับนักศึกษาคณะเทคโนโลยีสารสนเทศ. วารสาร เทคโนโลยีสารสนเทศ ลาดกระบัง, 4(2).

วิกิพีเดีย. (2556). การเรียนรู้แบบไม่มีผู้สอน. https://th.wikipedia.org/wiki/%E0%B8%81%E0%B8%B2%E0%B8%A3%E0%B9%80%E0%B8%A3%E0%B8%B5%E0%B8%A2%E0%B8%99%E0%B8%A3%E0%B8%B9%E0%B9%89%E0%B9%81%E0%B8%9A%E0%B8%9A%E0%B9%84%E0%B8%A1%E0%B9%88%E0%B8%A1%E0%B8%B5%E0%B8%9C%E0%B8%B9%E0%B9%89%E0%B8%AA%E0%B8%AD%E0%B8%99

วิกิพีเดีย. (2557). การเรียนรู้แบบมีผู้สอน. https://th.wikipedia.org/wiki/การเรียนรู้แบบมีผู้สอน

วิกิพีเดีย. (2562). การเรียนรู้ของเครื่อง. https://th.wikipedia.org/wiki/%E0%B8%81%E0%B8%B2%E0%B8%A3%E0%B9%80%E0%B8%A3%E0%B8%B5%E0%B8%A2%E0%B8%99%E0%B8%A3%E0%B8%B9%E0%B9%89%E0%B8%82%E0%B8%AD%E0%B8%87%E0%B9%80%E0%B8%84%E0%B8%A3%E0%B8%B7%E0%B9%88%E0%B8%AD%E0%B8%87

สมกันธา, ก., กุลตังวัฒนา, ว., หัสโก, ธ., และ รอดชมภู, จ. (2532). ระบบทำนายผลการเรียนนักศึกษาออนไลน์โดยใช้เคเนียเร็ซเนเบอะ (*Online Student Forecast System By Using K-Nearest Neighbor)*

ประวัติผู้เขียน

ชื่อ-สกุล	นางสาวขวัญตา ศิลปไพบูลย์พานิช
วัน เดือน ปี เกิด	1 กันยายน 2534
สถานที่เกิด	สุพรรณบุรี
วุฒิการศึกษา	พ.ศ. 2553 ครุศาสตรบัณฑิต สาขาวิชาคอมพิวเตอร์และเทคโนโลยีสารสนเทศ จาก มหาวิทยาลัยราชภัฏนครปฐม พ.ศ.2561 วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ จาก มหาวิทยาลัยศรีนครินทรวิโรฒ
ที่อยู่ปัจจุบัน	81/2 ม.2 ต.ตะค่า อ.บางปลาม้า จ.สุพรรณบุรี