การพยากรณ์ผลสัมฤทธิ์ทางการเรียน
ด้วยเทคนิคเหมืองข้อมูลโดยใช้ Rapid Miner

THE PREDICTION OF STUDENT PERFORMANCE
USING DATA MINING TECHNIQUES WITH RAPID MINER

จิราภรณ์ เจริญยิ่ง

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

2563

การพยากรณ์ผลสัมฤทธิ์ทางการเรียน
ด้วยเทคนิคเหมืองข้อมูลโดยใช้ Rapid Miner

จิราภรณ์ เจริญยิ่ง

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2563

THE PREDICTION OF STUDENT PERFORMANCE USING DATA MINING TECHNIQUES WITH RAPID MINER

JIRAPORN JAREANYING

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

2020

สารนิพนธ์

เรื่อง

การพยากรณ์ผลสัมฤทธิ์ทางการเรียน

ด้วยเทคนิคเหมืองข้อมูลโดยใช้ Rapid Miner

ของ

จิราภรณ์ เจริญยิ่ง

ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

ของมหาวิทยาลัยศรีนครินทรวิโรฒ

(รองศาสตราจารย์ นายแพทย์ฉัตรชัย เอกปัญญาสกุล)

คณบดีบัณฑิตวิทยาลัย

คณะกรรมการสอบปากเปล่าสารนิพนธ์

..... ที่ปรึกษาหลัก

(ผู้ช่วยศาสตราจารย์ ดร.วีรยุทธ เจริญเรืองกิจ)

..... ประธาน

(อาจารย์ ดร.รัตน์ชัยนันท์ ธรรมสุจริต)

..... กรรมการ

(อาจารย์ ดร.โสภณ มงคลลักษมี)

ชื่อเรื่อง	การพยากรณ์ผลสัมฤทธิ์ทางการเรียน ด้วยเทคนิคเหมืองข้อมูลโดยใช้ Rapid Miner
ผู้วิจัย	จิราภรณ์ เจริญยิ่ง
ปริญญา	วิทยาศาสตรมหาบัณฑิต
ปีการศึกษา	2563
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร. วีรยุทธ เจริญเรืองกิจ

งานวิจัยนี้นำเสนอการทำเหมืองข้อมูลทางการศึกษา โดยการจำแนกประเภทข้อมูล การวิเคราะห์องค์ประกอบหลักของข้อมูล เพื่อหาความสัมพันธ์ของตัวแปร และเปรียบเทียบประสิทธิภาพของอัลกอริทึม ซึ่งเป็นการศึกษาด้วยเทคนิค ต้นไม้ตัดสินใจ เทคนิคป่าแห่งการทำนาย การเรียนรู้เบย์ และ K-NN โดยใช้ข้อมูลของนักเรียนระดับมัธยมศึกษาในโรงเรียนโปรตุเกส ประกอบด้วยข้อมูลด้านผลการเรียน ด้านความเป็นอยู่ และความเชื่อมโยงทางสังคมและโรงเรียน จาก UCI Machine Learning Repository มีข้อมูล 649 รายการ 31 แอตทริบิวต์ จากผลการวิจัยพบว่าเทคนิคที่ให้ค่าความถูกต้อง มากที่สุด คือเทคนิคป่าไม้ตัดสินใจ เท่ากับ 80.74 และเทคนิคที่มีค่าความถ่วงดุลมากที่สุด คือ เทคนิคการเรียนรู้เบย์ เท่ากับ 55.52

คำสำคัญ : เทคนิคเหมืองข้อมูล, การจำแนกประเภทข้อมูล, การพยากรณ์ผลการเรียน, เทคนิคต้นไม้ตัดสินใจ, เทคนิคป่าเพื่อการทำนาย, เทคนิคการเรียนรู้เบย์, เทคนิคเพื่อนบ้านใกล้ที่สุด, โปรแกรม Rapid Miner

Title	THE PREDICTION OF STUDENT PERFORMANCE USING DATA MINING TECHNIQUES WITH RAPID MINER
Author	JIRAPORN JAREANYING
Degree	MASTER OF SCIENCE
Academic Year	2020
Thesis Advisor	Assistant Professor Dr. Werayuth Charoenruengkit

This research presents education data mining using data classification and Principal Component Analysis (PCA) to predict student grades using several variables and algorithms. The algorithms used in this paper are Decision Tree, Random Forest, Naïve-Bayes, and K-NN (K-Nearest Neighbors). The dataset experiment included secondary school students in Portuguese schools containing information about the socioeconomic environment of the students and learning environment from the UCI Machine Learning Repository. The population consisted of 649 samples with 31 attributes. The experimental results showed that the best accuracy from the Random Forest technique was 80.74 and the Naïve-Bayes technique had the highest average at 55.52.

Keyword : Data Mining, Classification, Student Performance Prediction, Decision Tree, Random Forest, Naïve-Bayes, k-Nearest Neighbor, Rapid Miner

กิตติกรรมประกาศ

สารนิพนธ์เล่มนี้สำเร็จสมบูรณ์ได้ด้วยดีเพราะได้รับความกรุณาชี้แนะและช่วยเหลืออย่างดียิ่งจาก ผู้ช่วยศาสตราจารย์ ดร. วีรยุทธ เจริญเรืองกิจ อาจารย์ที่ปรึกษาสารนิพนธ์ ที่ให้คำแนะนำและตรวจแก้ไขข้อบกพร่องมาโดยตลอด ตั้งแต่เริ่มต้นจนสำเร็จเรียบร้อย และคณะอาจารย์ทุกท่านที่ให้ความรู้ในสาระวิชาต่าง ๆ ในการศึกษาระดับปริญญาโทมาตลอดการศึกษานี้ผู้วิจัยขอกราบขอบพระคุณด้วยความเคารพอย่างสูง

สุดท้ายนี้ขอขอบคุณบิดา มารดา และครอบครัว ที่ให้การสนับสนุนและให้กำลังใจมาให้โดยตลอด และขอบคุณเพื่อน ๆ พี่ ๆ และน้อง ๆ ที่ร่วมเรียนสาขาเทคโนโลยีสารสนเทศซึ่งได้ให้ความช่วยเหลือทั้งทางตรงและทางอ้อมไว้ ณ โอกาสนี้

จิราภรณ์ เจริญยิ่ง

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

1.1 ความสำคัญและที่มาของงานวิจัย

การศึกษาเป็นสิ่งสำคัญที่ช่วยในการพัฒนาบุคคลให้มีความรู้ความสามารถในด้านต่าง ๆ รวมทั้งทัศนคติและพฤติกรรมอื่น ๆ ซึ่งเป็นฐานสำคัญให้บุคคลนั้นมีความเจริญงอกงามทางปัญญา ดำรงชีวิตได้อย่างเหมาะสม ส่งผลต่อการพัฒนาประเทศในลำดับต่อไป ทั้งในด้านสังคม เศรษฐกิจ และการเมือง

การวัดและประเมินผลการเรียนรู้ของผู้เรียนเป็นองค์ประกอบสำคัญในการพัฒนาคุณภาพการศึกษา ทำให้ได้ข้อมูลสารสนเทศที่จำเป็นในการพิจารณาว่าผู้เรียนเกิดคุณภาพการเรียนรู้ตามผลการเรียนรู้ที่คาดหวังและมาตรฐานการเรียนรู้ การวัดและประเมินผลทางการศึกษาจะเกี่ยวข้องกับคำ 3 คำ คือ (สมชาย รัตนทองคำ, 2554)

1. การทดสอบ (testing) หมายถึง การนำเสนอชุดคำถามที่เรียกว่าข้อสอบหรือแบบทดสอบที่มีมาตรฐานให้ผู้สอบตอบ

2. การวัดผล (measurement) หมายถึง การวัดคุณลักษณะ (ตัวแปร)ของบุคคลจากผลการตอบคำถามในแบบทสอบตามกฎเกณฑ์ที่กำหนด เพื่อแสดงคุณค่าเชิงปริมาณหรือตัวเลขที่วัดได้ การวัดผลนอกจากใช้แบบทดสอบแล้วยังรวมถึงการใช้เครื่องมืออื่นเพื่อรวบรวมข้อมูลเชิงปริมาณหรือเชิงคุณภาพด้วย เช่น การสังเกตพฤติกรรม การสัมภาษณ์การตรวจผลงานต่าง ๆ ที่กำหนดให้ผู้ประเมินทำ

3. การประเมินผล (evaluation) หมายถึง กระบวนการอย่างมีระบบที่นำข้อมูลจากการวัดผลมาตีค่าและตัดสินคุณค่าของผู้เรียน ซึ่งการวัดผลและการประเมินผลเป็นกระบวนการที่มีความต่อเนื่อง เมื่อมีการวัดผลจะทำให้ได้ข้อมูลและรายละเอียดหลายด้าน เมื่อนำข้อมูลดังกล่าวมาวิเคราะห์เปรียบเทียบกับเกณฑ์ใด เกณฑ์หนึ่งเพื่อตีค่า หรือสรุปคุณค่าออกมาถือว่าเป็นกระบวนการประเมิน ผลการประเมินจะมีความถูกต้องเที่ยงตรงเพียงใดขึ้นกับความถูกต้องของผลการวัด ถ้าผลการวัดถูกต้องการประเมินก็จะมีความเชื่อถือได้มากและตรงกับความเป็นจริง ถ้าผลการวัดผิดพลาด การประเมินก็จะผิดพลาดไปด้วย การวัดผลและการประเมินผลมีความแตกต่างกัน

โดยการวัดและประเมินผลการเรียนรู้ของผู้เรียนนั้น มีจุดประสงค์เพื่อพัฒนาผู้เรียน แต่ข้อมูลการวัดและประเมินผลการเรียนรู้เป็นสิ่งที่ครูผู้สอนยังไม่ทราบข้อมูลได้ในตอนต้น ดังนั้นจึง

ต้องนำข้อมูลอื่น ๆ ที่เกี่ยวข้องกับตัวผู้เรียนมาใช้ในการทำนาย เพื่อให้ทราบผลการประเมินจึงจะสามารถส่งเสริมหรือแก้ไขการเรียนรู้รวมทั้งการสอนของครูให้มีคุณภาพมากยิ่งขึ้น

จากที่มาและความสำคัญดังกล่าว ผู้วิจัยได้เห็นความสำคัญของประเมินผลสัมฤทธิ์ทางการเรียน จึงมีแนวคิดที่จะใช้ข้อมูลของผู้เรียนมาวิเคราะห์ด้วยเทคนิคเมืองข้อมูลในการทำนายผลสัมฤทธิ์ทางการเรียน และได้เลือกทำการศึกษาข้อมูลตัวอย่าง ซึ่งเป็นข้อมูลของนักเรียนระดับมัธยมศึกษาในโรงเรียนโปรตุเกส ประกอบด้วยข้อมูลด้านผลการเรียน ด้านความเป็นอยู่ และความเชื่อมโยงทางสังคมและโรงเรียน จาก UCI Machine Learning Repository (Cortez, 2008) มีข้อมูล 649 รายการ 33 แอตทริบิวต์ เพื่อนำผลการพยากรณ์ที่ได้มาช่วยให้คุณครูได้ทราบถึงกลุ่มเด็กที่ต้องให้ความสนใจเป็นพิเศษ เพื่อคอยสนับสนุนและให้ความช่วยเหลือ จนทำให้ผลการเรียนของผู้เรียนนั้นบรรลุตามวัตถุประสงค์ต่อไป

1.2 วัตถุประสงค์ของงานวิจัย

1.เพื่อศึกษาเทคนิคที่เหมาะสมในการพยากรณ์ผลสัมฤทธิ์ทางการเรียนด้วยการทำเหมืองข้อมูล

2.เพื่อพยากรณ์ผลสัมฤทธิ์ทางการเรียนในเทอมที่ 3 ของนักเรียนในข้อมูลตัวอย่าง

3.เพื่อสำรวจตัวแปรที่สามารถทำนายผลสัมฤทธิ์ทางการเรียนของนักเรียน

1.3 ขอบเขตของงานวิจัย

การวิจัยนี้เป็นการทำเหมืองข้อมูลทางการศึกษา (Education Data Mining) โดยการจำแนกประเภทข้อมูล (Classification) หาความสัมพันธ์ของข้อมูล และเปรียบเทียบประสิทธิภาพของอัลกอริทึม โดยเป็นการศึกษาด้วยเทคนิค ต้นไม้ตัดสินใจ (Decision Tree) เทคนิคป่าแห่งการทำนาย (Random Forest) การเรียนรู้เบย์ (Naïve-Bayes) และ K-NN (K-Nearest Neighbors) โดยใช้ซอฟต์แวร์ Rapid Miner Studio ในการประมวลผลข้อมูล และสร้างตัวแบบทำนายการพยากรณ์ผลสัมฤทธิ์ทางการเรียน

1.4 วิธีดำเนินงานวิจัย

1. ศึกษางานวิจัยที่เกี่ยวข้อง: Literature Review โดยศึกษาทฤษฎีและศึกษางานวิจัยที่เกี่ยวกับเทคนิคการทำเหมืองข้อมูล (Data Mining)

2. กำหนดขอบเขตการศึกษา เลือกเครื่องมือที่ใช้ในการประมวลผลการพยากรณ์

- ชุดข้อมูล คือ การเตรียมข้อมูล, วิเคราะห์ข้อมูลและค่าต่าง ๆ ที่จะทำนาย

- เครื่องมือ คือ ศึกษาเครื่องมือ และเลือกเครื่องมือที่จะนำมาวิเคราะห์ข้อมูล

3. ศึกษาทฤษฎี และเทคนิคต่าง ๆ ในการทำเหมืองข้อมูล (Data Mining)

4. ใช้ซอฟต์แวร์ RapidMiner Studio เพื่อทำการวิเคราะห์ข้อมูล ด้วยเทคนิค (Decision Tree) เทคนิคป่าแห่งการทำนาย (Random Forest) การเรียนรู้เบย์ (Naïve-Bayes) และ K-NN (K-Nearest Neighbors) และตรวจสอบความถูกต้อง

5. ทำการสรุป และอภิปรายผลของงานวิจัย

1.5 ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัย

1. สามารถศึกษาการจัดการข้อมูลและการจำแนกประเภทของข้อมูลได้

2. สามารถศึกษาการวิเคราะห์องค์ประกอบหลักของข้อมูลได้ (Principal Component Analysis

3. สามารถศึกษาการใช้เทคนิค (Decision Tree) เทคนิคป่าแห่งการทำนาย (Random Forest) การเรียนรู้เบย์ (Naïve-Bayes) และ K-NN (K-Nearest Neighbors) ในการสร้างตัวแบบทำนายเพื่อการพยากรณ์ผลสัมฤทธิ์ทางการเรียนได้

4. สามารถศึกษาการใช้ Rapid Miner Studio ซึ่งเป็นเครื่องมือวิเคราะห์ข้อมูลและสร้างตัวแบบทำนายการพยากรณ์ผลสัมฤทธิ์ทางการเรียนได้

บทที่ 2
แนวคิด ทฤษฎี เทคโนโลยีที่เกี่ยวข้อง

2.1 ทฤษฎีพื้นฐานเกี่ยวกับการทำเหมืองข้อมูล (Data Mining)

Data Mining เป็นกระบวนการ (Process) ที่กระทำกับข้อมูลขนาดใหญ่ เพื่อค้นหารูปแบบ แนวทาง และความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น โดยอาศัยหลักสถิติการรู้จากการเรียนรู้ของเครื่องและหลักคณิตศาสตร์ ทำการจำแนกประเภท รูปแบบ เชื่อมโยงข้อมูลที่มีความสัมพันธ์กัน และหาความน่าจะเป็นที่จะเกิดขึ้น เพื่อให้ได้องค์ความรู้ใหม่ ที่สามารถนำไปใช้ประกอบการตัดสินใจในด้านต่าง ๆ เพื่อให้ได้สารสนเทศที่เราไม่รู้ออกมา โดยสารสนเทศที่ได้จะมีเหตุผลสนับสนุนและสามารถนำไปใช้ประโยชน์ได้ (อดุลย์ ยิ้มงาม, 2551) ซึ่งมาตรฐานที่ใช้สำหรับการทำเหมืองข้อมูล เพื่อทำการวิเคราะห์และนำไปใช้ประโยชน์ ดังภาพประกอบที่ 1 คือ Cross-industry standard process for data mining : CRISP-DM ประกอบด้วย 6 ขั้นตอน คือ (Data Cube, 2564)

ภาพประกอบ 1 ขั้นตอนในการทำ CRISP-DM

ที่มา: https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

1. การทำความเข้าใจธุรกิจ (Business Understanding) เป็นการทำความเข้าใจข้อมูล ปัญหาและวัตถุประสงค์ จากนั้นแปลงปัญหาให้อยู่ในรูปของโจทย์สำหรับการวิเคราะห์ข้อมูล และวางแผนการดำเนินงานเบื้องต้น

2. การทำความเข้าใจข้อมูล (Data Understanding) เป็นการรวบรวมข้อมูลที่ทำความเข้าใจ จากนั้นตรวจสอบคุณภาพ และเลือกข้อมูลที่เก็บรวบรวมมาว่าจะใช้ข้อมูลใดบ้างในการวิเคราะห์

3. การเตรียมข้อมูล (Data Preparation) เป็นการเตรียมข้อมูลทั้งหมดที่จะทำ เพื่อให้ข้อมูลดิบที่เรารวบรวมมากลายเป็นข้อมูลสมบูรณ์ที่พร้อมจะเข้าสู่ตัวแบบทำนาย เช่น การสร้างตาราง การลบข้อมูลที่ไม่ต้องการออก การแปลงข้อมูลให้อยู่ในรูปแบบที่ต้องการ เป็นต้น

4. การสร้างตัวแบบทำนาย (Modeling) เป็นขั้นตอนที่จะเลือกและทดสองสร้างตัวแบบทำนายหลาย ๆ แบบที่น่าจะสามารถแก้ไขปัญหาที่ต้องการได้ จากนั้นค่อย ๆ ปรับค่าพารามิเตอร์ในแต่ละตัวแบบทำนาย เพื่อให้ได้ตัวแบบทำนายที่เหมาะสมที่สุดมาใช้ในการแก้ไขปัญหา หากยังไม่ได้ตัวแบบทำนายที่พอใจ ก็สามารถกลับไปเตรียมข้อมูลให้พร้อมมากกว่านี้ได้ เนื่องจากข้อมูลที่ดี ก็จะทำให้ได้ผลลัพธ์ที่ดีเช่นกัน

5. การวัดประสิทธิภาพของตัวแบบทำนาย (Evaluation) เป็นการวัดประสิทธิภาพของตัวแบบทำนายที่ได้จากขั้นตอนที่ 4 เพื่อวัดว่าตัวแบบทำนายมีประสิทธิภาพเพียงพอต่อการนำไปใช้งานแล้วหรือไม่ ซึ่งตัวแบบทำนายแต่ละประเภทก็จะมีตัววัดประสิทธิภาพที่แตกต่างกันออกไป

6. การนำตัวแบบทำนายไปใช้งานจริง (Deployment) เป็นการนำตัวแบบทำนายที่เหมาะสมที่สุดไปใช้งานจริง เพื่อวิเคราะห์และแก้ปัญหาที่ต้องการ

2.2 การแปลงข้อความเป็นตัวเลข (Label Encoding)

เป็นการแปลงข้อมูลประเภทข้อความให้เป็นตัวเลข 0 1 2 3 โดยทำให้ข้อมูลในแต่ละคอลัมน์เป็น id ที่ต่างกัน แต่ยังคงอยู่ในคอลัมน์เดิม ซึ่งจะทำการแปลงข้อมูลเมื่อข้อมูลเป็นข้อมูลที่เป็นอันดับหรือข้อมูลที่ไม่ได้เป็นตัวเลข ซึ่งข้อมูลที่ Encode แบบ Label Encoding จะช่วยทำให้โมเดลใช้พื้นที่ในการเก็บข้อมูลน้อยลง และลดเวลาในการคำนวณด้วย จากตารางที่ 1 สัตว์เลี้ยงของนาย A นาย B นาย C และ นาย D (ชิตพงษ์ กิตตินราดร, 2562)

ตาราง 1 ตารางแสดงสัตว์เลี้ยง

ชื่อ	สัตว์เลี้ยง
คุณ A	Dog
คุณ B	Cat
คุณ C	Dog
คุณ D	Mouse

เมื่อเราแปลงข้อมูล Column สัตว์เลี้ยง เป็น Label Encoder แล้ว จะได้ดังตารางที่ 2

ตาราง 2 ตารางแสดงสัตว์เลี้ยงหลังจากแปลงเป็น Label Encoder

ชื่อ	สัตว์เลี้ยง
คุณ A	1
คุณ B	2
คุณ C	1
คุณ D	3

2.3 การแปลงข้อความเป็นตัวเลขฐาน 2 (One-hot Encoding)

เป็นการแปลงข้อมูลประเภทข้อความให้แตกเป็นคอลัมน์ย่อย ๆ แบบ Binary 0/1 ตามค่าของข้อมูล โดยใช้ 0,1 แทนว่า item นั้นอยู่ในคอลัมน์ใดและข้อมูลยังคงมีความสัมพันธ์กันอยู่ ซึ่งข้อมูลที่ Encode แบบ One Hot จะช่วยให้โมเดล Machine Learning ทำงานง่ายขึ้น แทนที่โมเดลจะเรียนรู้ Pattern จากสัญญาณเดียว 3 ระดับ เปลี่ยนมาเรียนรู้จากสวิทช์ปิดเปิด 3 ตัวแทน ความหมายของข้อมูลแบบ Nominal จะตรงขึ้น เนื่องจาก สีแดง 1 ไม่ได้ใกล้เคียง สีเขียว 2 มากกว่าสีน้ำเงิน 3 จากตารางที่ 3 ใน Column สีเสื้อ แดง = 1, เขียว = 2 และ น้ำเงิน = 3 (ชิตพงษ์ กิตตินราดร, 2562)

ตาราง 3 ตารางแสดงสีเสื้อ แดง 1 เขียว 2 น้ำเงิน 3

ชื่อ	จำนวนบุตร	สีเสื้อ
คุณ A	4	2
คุณ B	3	1
คุณ C	1	3
คุณ D	2	1

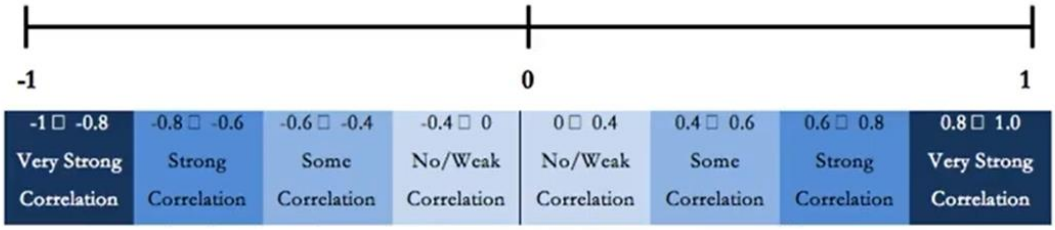
เมื่อเราแปลงข้อมูล Column สีเสื้อ เป็น One Hot Encoding แล้ว จะได้ดังตารางที่ 4

ตาราง 4 ตารางสีเสื้อหลังจากแปลงเป็น One Hot Encoding

ชื่อ	จำนวนบุตร	สีเสื้อแดง	สีเสื้อเขียว	สีเสื้อน้ำเงิน
คุณ A	4	0	1	0
คุณ B	3	1	0	0
คุณ C	1	0	0	1
คุณ D	2	1	0	0

2.4 การหาค่าสหสัมพันธ์ (Correlation)

เป็นการศึกษาความสัมพันธ์ระหว่างตัวแปรตั้งแต่ 2 ตัวขึ้นไป เช่น การหาความสัมพันธ์ระหว่างอายุกับน้ำหนัก อายุกับรายได้ ซึ่งใช้ค่า Correlation เป็นค่าวัดความสัมพันธ์ โดยใช้วิธีการทางสถิติหลากหลายวิธี การจะใช้ค่าสถิติตัวใดขึ้นอยู่กับลักษณะของตัวแปร และจะต้องมีการทดสอบนัยสำคัญก่อน จึงจะสรุปว่าตัวแปรนั้น ๆ สัมพันธ์กันหรือไม่ การอ่านผลจะอ่านเพียงว่าทั้งสองตัวมีความสัมพันธ์กันหรือไม่ ถ้าสัมพันธ์กัน แปรผันตามกันหรือแปรผกผันกันเป็นค่ามากน้อยเพียงใด แต่จะไม่สามารถบอกว่าตัวแปรใดเป็นตัวแปรต้น หรือตัวแปรใดเป็นตัวตาม



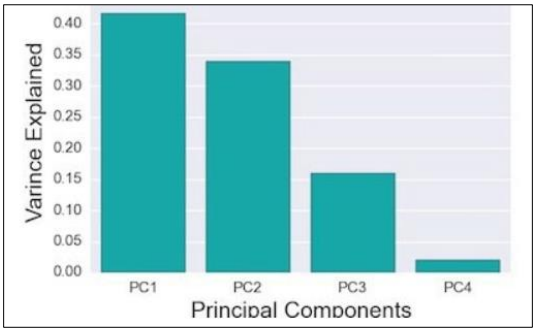
ภาพประกอบ 2 ระดับค่าความสัมพันธ์ของ Correlation

ที่มา : https://jupiter.ba.cmu.ac.th/data-mining-with-rapidminer-สำหรับผู้เริ่มต้น/

จากภาพประกอบที่ 2 จะแสดงระดับค่าความสัมพันธ์ของ Correlation โดยค่า Correlation จะมีค่าระหว่าง -1 ถึง 1 ซึ่งระดับความสัมพันธ์ตั้งแต่ 0 – 0.4 และ 0 – (-4) หมายถึง ตัวแปรมีสัมพันธ์ไม่มีความสัมพันธ์กันเลย หรือมีความสัมพันธ์กันน้อยมาก 0.41 - 0.6 และ (-0.41) – (-0.6) หมายถึง ตัวแปรมีสัมพันธ์กันบ้าง 0.61 – 0.8 และ (-0.61) – (-0.8) หมายถึง ตัวแปรมีสัมพันธ์กันมาก 0.81 – 1.0 และ (-0.81) – (-1.0) หมายถึง ตัวแปรมีสัมพันธ์กันมากที่สุด

2.5 การวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis : PCA)

PCA (ธีระดา ภิญโญ, 2561) เป็นการวิเคราะห์เพื่อหาความสัมพันธ์ของข้อมูลที่มีหลายตัวแปร หลักการนี้จะไม่ส่งผลกระทบต่อข้อมูลหลัก เพียงแค่ปรับเปลี่ยนมุมมองข้อมูลใหม่ ให้ข้อมูลมีความกระชับ มีขนาดเล็กลงง่ายต่อการนำไปใช้งาน ซึ่งทำให้ข้อมูลที่มีความซับซ้อน เกิดการลดขนาดของ Matrix มีขนาดเล็กลงและง่ายต่อการอธิบาย และช่วยให้ใช้เวลาการสร้างตัวแบบทำนายน้อยลง เช่น มีตัวแปร 50 ตัว แต่อาจจะไม่ได้ใช้งานทั้งหมด โดยจะเรียงลำดับความสำคัญของตัวแปร และพิจารณาว่าตัวแปรใดส่งผลกระทบก็จะนำมาใช้ ส่วนตัวแปรใดที่ไม่มีความเกี่ยวข้องก็จะไม่ได้นำมาสร้างโมเดล



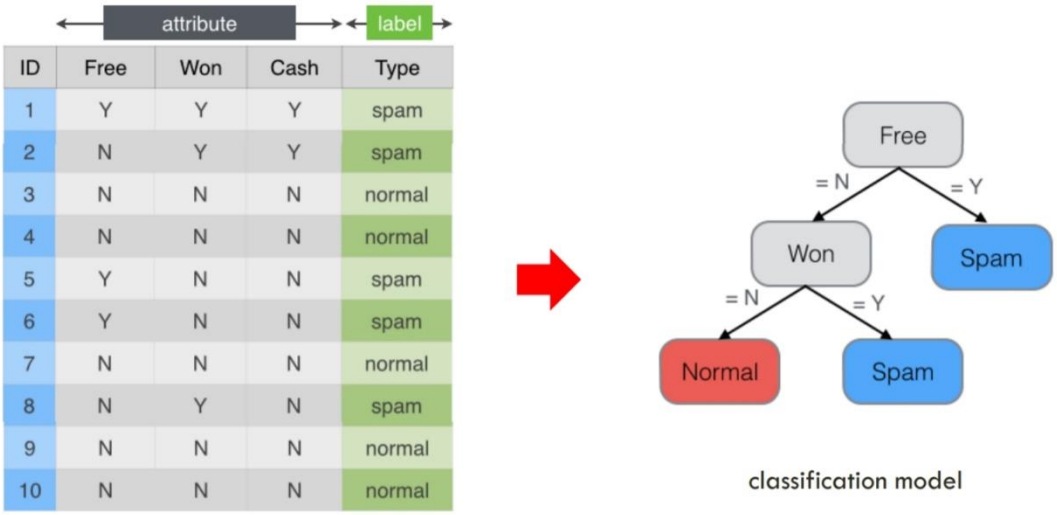
ภาพประกอบ 3 แบบจำลองการสร้าง PCA

ที่มา : https://kongruksiamza.medium.com/สรุป-machine-learning-ep-5-การวิเคราะห์องค์ประกอบหลัก-pca-185c29aa8954

จากภาพประกอบที่ 3 จะเห็นว่า PCA จะทำการสร้างตัวแปรที่เรียกว่า component โดยแต่ละ component จะไม่มีความสัมพันธ์กันเลย component ตัวแรกจะมีค่า variance สูงที่สุด ซึ่งอธิบายข้อมูลได้มากที่สุด และตัวถัดๆ ไปก็จะมี variance ลดลงตามลำดับ โดยค่าที่ครอบคลุมจะมี variance ประมาณ 80–90% แต่หลังจากใช้ PCA บนชุดข้อมูลแล้ว คุณลักษณะดั้งเดิมของข้อมูลจะเปลี่ยนไปเป็นส่วนประกอบหลัก ซึ่งส่วนประกอบหลักนั้นจะไม่สามารถอ่านและตีความได้เหมือนกับคุณสมบัติดั้งเดิม

2.5 เทคนิคต้นไม้ตัดสินใจ (Decision Tree)

เป็นการเรียนรู้โดยการจำแนกประเภทข้อมูลออกเป็นกลุ่ม โดยใช้คุณสมบัติของข้อมูล เป็นตัวกำหนด ซึ่งประกอบไปด้วย โหนดภายใน, กิ่ง และโหนดใบ วิธีการวิเคราะห์แบบต้นไม้ตัดสินใจเป็นการค้นหาจากบนลงล่าง โดยเริ่มจากการเลือกคุณสมบัติที่ดีที่สุดมาเป็นโหนดราก และวนสร้างโหนดลูกและเส้นเชื่อมไปเรื่อย ๆ จนกว่าข้อมูลที่ได้จะถูกจัดไว้เป็นกลุ่มเดียวกันถึงจะหยุดสร้างต้นไม้ จากการคำนวณ Information Gain (IG) โดยเลือกคุณสมบัติที่มีค่า IG สูงที่สุด หรือการคำนวณค่า Gini Index ซึ่งเป็นค่าที่บ่งบอกว่า ตัวแปรใดสมควรนำมาใช้ในการแบ่ง และคำนวณที่หาค่า Gini Index ที่น้อยที่สุดจะคือค่าที่ดีที่สุด จะนำตัวแปรนั้นมาเป็นโหนดราก (ชณิดาภา บุญประสม และ จรัญ แสนราช, 2561)



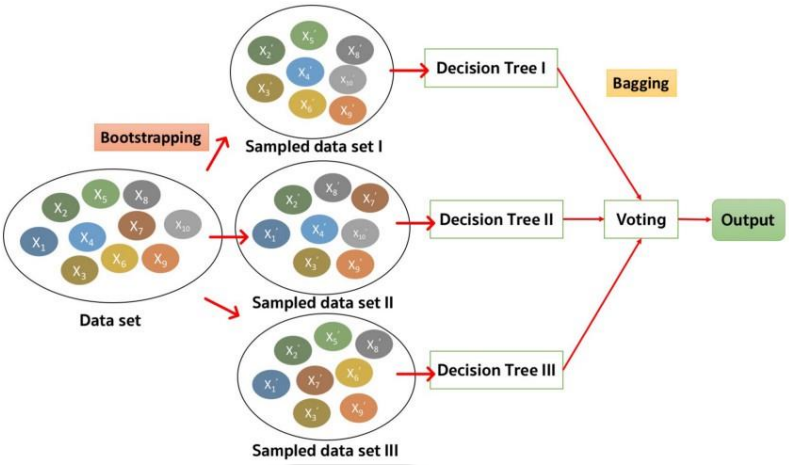
ภาพประกอบ 4 การสร้างโมเดลด้วยเทคนิคต้นไม้ตัดสินใจ

ที่มา : https://dataminingtrend.com/2014/data-minimg-techniques/ensemble-model/

จากภาพประกอบที่ 4 เป็นการทำนายอีเมลเป็น spam mail หรือไม่ โดยการคำนวณค่า Information Gain (IG) ของแต่ละแอตทริบิวท์ คือ free won และ cash หากแอตทริบิวท์ใดมีค่า IG สูงที่สุดก็จะนำมาเป็นโหนดราก และแอตทริบิวท์ที่มีค่ารองลงมาก็จะเป็นโหนดใบ และต่อกันไปเรื่อย ๆ จนสามารถบอกได้ว่าอีเมลฉบับนี้เป็น spam หรือ normal

2.6 **เทคนิคป่าแห่งการทำนาย** (Random Forest)

เป็นการเทรนตัวแบบทำนายที่เหมือนกันหลายๆ ครั้ง บนข้อมูลชุดเดียวกัน โดยแต่ละครั้งของการเทรนจะเลือกส่วนของข้อมูลที่เทรนไม่เหมือนกัน จากนั้นเอาการตัดสินใจของตัวแบบทำนายเหล่านั้นมาโหวตกันว่า Class ไหนถูกเลือกมากที่สุด ซึ่งจะช่วยแก้ปัญหา high variance ที่เกิดจาก tree แต่ละต้นได้ ซึ่งจะใช้ตัวแบบทำนายจากต้นไม้ตัดสินใจหลาย ๆ ต้น ตั้งแต่ 10 ถึงมากกว่า 1000 ตัวแบบทำนาย ซึ่งแต่ละตัวแบบทำนายจะได้รับชุดข้อมูลที่ไม่เหมือนกัน และจะทำนายแยกกันออกมา หลังจากนั้นจะเลือกผลการทำนายสุดท้าย จากค่าการทำนายที่ได้รับการโหวตมากที่สุด ดังภาพประกอบที่ 5 (ปรเมษฐ์ ธันวานนท์, ชัยกรยิ่ง เสรี, วรพล พงษ์เพ็ชร และ ธนภัทร ฆังคะจิตรสกสรรค์, 2560)



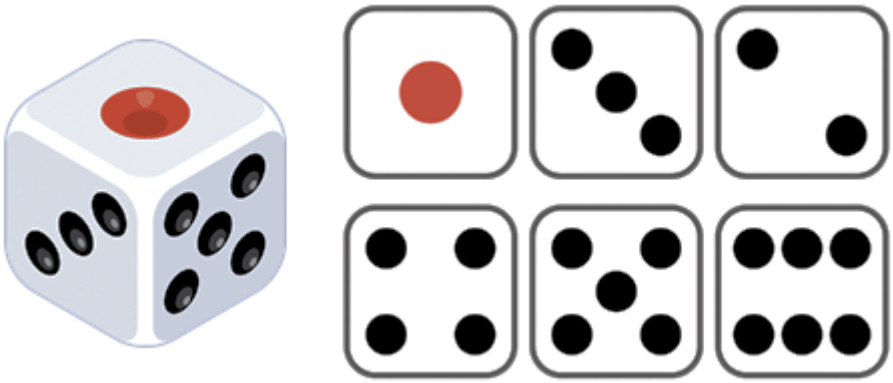
ภาพประกอบ 5 การสร้างป่าแห่งการทำนาย

ที่มา : https://medium.com/@witchapongdaroontham/เจาะลึก-random-forest-part-2-of-รู้จัก-decision-tree-random-forest-และ-xgboost-79b9f41a1c1c

ซึ่งค่าที่สามารถปรับได้ มีดังนี้ 1) number of tree คือ จำนวน tree ที่มีผลต่อประสิทธิภาพของ model โดยดูจากจุดที่ performance เริ่มจะนิ่ง จนจำนวน tree ไม่มีผลต่อ model performance แล้ว 2) criterion คือ cost function ที่เราจะใช้ สำหรับ classification tree นั้นจะ มีค่า default เป็น gini ซึ่งสามารถจะเลือกเป็น entropy ก็ได้ 3) max_depth คือ จำนวน level ที่มากที่สุด ของ node ที่จะทำการ split observation ซึ่งเราจะทำกำหนดค่า max_depth ไม่ให้มากเกินไป เพื่อป้องกันปัญหา overfitting

2.6 เทคนิคการเรียนรู้เบย์ (Naïve-Bayes)

เป็นตัวแบบทำนายการจำแนกประเภทข้อมูล โดยวิเคราะห์หาความน่าจะเป็นของสิ่งที่ยังไม่เคยเกิดขึ้น โดยการคาดเดาจากสิ่งที่เคยเกิดขึ้นมาก่อน ความน่าจะเป็นที่เกิดเหตุการณ์หนึ่ง ก็ต่อเมื่อเหตุการนึงได้เกิดไปแล้ว กล่าวคือเราสนใจจะหาความน่าจะเป็นที่จะเกิดเหตุการณ์ y ถ้ามีเหตุการ์ x เกิดขึ้นเล้ว โดยมีสมมติฐานว่าปริมาณของความสนใจขึ้นอยู่กับการกระจายความน่าจะเป็น(Probability Distribution) เช่น ความน่าจะเป็นคนที่มีงานแล้ว ได้อนุมัติเงินกู้ หรือโอกาสคนที่ปลอดหนี้บ้านจะได้อนุมัติเงินกู้เป็นเท่าใด ยิ่งมีตัวแปรในการพิจารณาจำนวนมาก การพิจารณาความน่าจะเป็นก็จะมากขึ้นตามไปด้วย (ชณิดาภา บุญประสม และ จรัญ แสนราช, 2561)



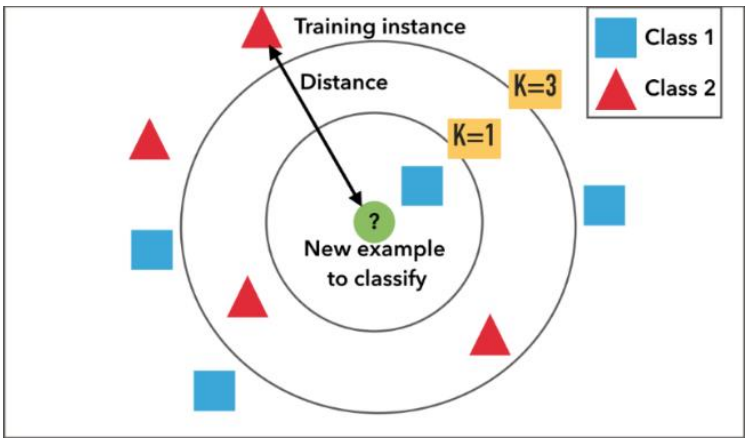
ภาพประกอบ 6 การทอยลูกเต๋าเพื่อคำนวณความน่าจะเป็น

ที่มา : http://cakeknowledgeblogs.blogspot.com/2019/08/naive-bayes-classification-1.html

จากภาพประกอบ 6 คือการทอยลูกเต๋า ซึ่งถ้าสมมุติโจทย์เราคือต้องการหาความน่าจะเป็นของการทอยลูกเต๋าหนึ่งลูกแล้วโอกาสที่จะได้เลขจำนวนคู่มีความน่าจะเป็นเท่าไหร่ ซึ่งสามารถรู้ได้ว่ามีโอกาสเกิดได้ 50 เปอร์เซ็นดังนี้ P(dice lands on an even number) = 3/6 = 1/2 = 0.5 จากสมการซึ่งรู้อยู่แล้วว่าลูกเต๋า 1 ลูกมีค่าทั้งหมด 6 ค่าคือ 1,2,3,4,5,6 ซึ่งเป็นเลขคู่คือเลข 2,4,6 ซึ่งมีทั้งหมด 3 ค่า ดังนั้นความน่าจะเป็นที่ออกได้เลขคู่จึงเกิดจาก 3/6=1/2=0.5

2.7 เทคนิคเพื่อนบ้านใกล้ที่สุด KNN (K-Nearest Neighbors)

เป็นวิธีการค้นหาเพื่อนบ้านใกล้ที่สุด K-Nearest Neighbors เป็นการแบ่งกลุ่มข้อมูล และทำการวัดระยะห่างระหว่างข้อมูลที่ต้องการทำนายกับข้อมูลที่อยู่ใกล้เคียงเป็นจำนวน K ตัว โดยการวัดระยะมี 2 แบบ คือ ฟังก์ชันระยะทาง(Distance Function) เป็นการคำนวณค่าระยะห่างระหว่างสองเรคคอร์ด เพื่อที่จะมาวัดความคล้ายคลึงกันของข้อมูล และฟังก์ชันการแจกแจง (Combination Function) เป็นการรวมกันของผลลัพธ์ที่ได้จากการคำนวณค่าระยะห่าง(Distance) โดยทำการเรียงลำดับค่าระยะห่าง(Distance) จากน้อยไปมาก หลังจากนั้นดูจากค่า “K” ว่ากำหนดเป็นเท่าไร แล้วนำลำดับที่เรียงได้มาเทียบกับคลาสข้อมูลที่เรียงแล้วนำมาตอบ โดยทั้ง 2 แบบเหมาะสำหรับข้อมูลแบบตัวเลขเพื่อหาวิธีการวัดระยะห่างของแต่ละ Attribute ในข้อมูลได้ (ชณิดาภา บุญประสม และ จรัญ แสนราช, 2561)



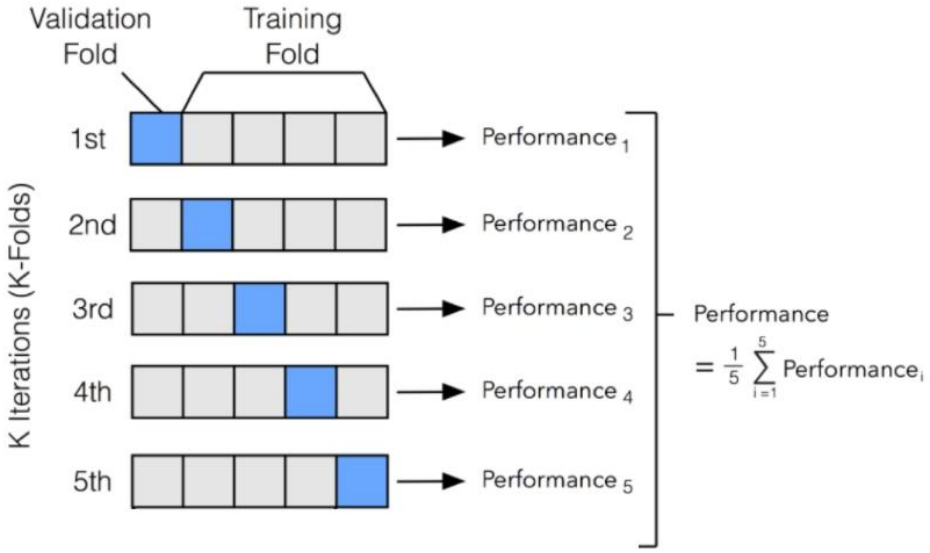
ภาพประกอบ 7 การแบ่งกลุ่มข้อมูลด้วยเทคนิคเพื่อนบ้านใกล้ที่สุด

ที่มา : https://kongruksiamza.medium.com/สรุป-machine-learning-ep-4-เพื่อนบ้านใกล้ที่สุด-k-nearest-neighbors-787665f7c09d

จากภาพประกอบที่ 7 เป็นการหาว่าจุดสีเขียวเป็น class 1 หรือ class 2 ด้วยการกำหนดค่า k และเมื่อกำหนด k = 1 จะทำนายว่าเป็นคลาส 1 เพราะในวงกลมจะมีแค่สี่เหลี่ยมสีฟ้า 1 รูป แต่เมื่อกำหนด k = 3 จะทำนายว่าเป็นคลาส 2 เพราะในวงกลมจะมีสี่เหลี่ยมสีฟ้า 1 รูป และมีสามเหลี่ยมสีแดง 2 รูป

2.8 การวิเคราะห์ความแม่นตรงของตัวแบบทำนาย K-fold Cross Validation

เป็นการวิเคราะห์ความแม่นตรงของตัวแบบทำนาย K-fold cross validation วิธีการตรวจสอบค่าความผิดพลาดในการคาดการของตัวแบบทำนาย โดยการแบ่งข้อมูลออกเป็นกลุ่มจำนวน K กลุ่ม (K-fold) ในตอนแรกเลือกข้อมูลกลุ่มที่ 1 เป็นข้อมูลชุดทดสอบ และข้อมูลชุดที่เหลือจะเป็นข้อมูลชุดสอน นำข้อมูลไปจัดหมวดหมู่ จากนั้นจะสลับข้อมูลกลุ่มที่ 2 มาเป็นชุดทดสอบและข้อมูลกลุ่มอื่น ๆ ที่เหลือ เป็นชุดทดสอบ สลับอย่างนี้ไปเรื่อย ๆ จนครบ K กลุ่ม ในขั้นตอนสุดท้ายจะหาค่าเฉลี่ยและค่าส่วนเบี่ยงเบนมาตรฐานของค่าความถูกต้องในแต่ละกลุ่ม โดยวิธีการนี้ข้อมูลทุกตัวจะได้เป็นทั้งชุดทดสอบและชุดสอน ซึ่งจะดีกว่าการแบ่งข้อมูลไว้เป็น 2 ส่วนคือ Training Data กับ Testing Data ที่ทำให้ไม่ทราบได้ว่าข้อมูลใดคือข้อมูลที่ควรจะนำมาเป็น Training Data แต่ Cross Validation จะสามารถหา Training Data ที่ดีที่สุด โดยการแบ่งข้อมูลและนำทุกส่วนมาเป็น Training Data และยังสามารถใช้เปรียบเทียบได้อีกว่าควรใช้วิธีไหนที่เหมาะสมที่สุดในการสร้างโมเดล (รัชพล กลัดชื่น และ จรัญ แสนราช, 2561)



ภาพประกอบ 8 ขั้นตอนการวิเคราะห์ความแม่นตรงของตัวแบบทำนาย K-fold cross validation

ที่มา : https://www.kaggle.com/discussion/204878

จากภาพประกอบที่ 8 เป็นการวิเคราะห์ความแม่นตรงของตัวแบบทำนาย โดยกำหนดให้ k = 5 แล้วเรียงลำดับการทดสอบซึ่งเป็นข้อมูลในช่องสีฟ้าเป็นข้อมูลทดสอบ และข้อมูลในช่องสีเทาเป็นข้อมูลชุดสอน โดยสลับข้อมูลไปเรื่อย ๆ จนครบทุกช่อง

2.9 การวัดประสิทธิภาพของตัวแบบทำนาย

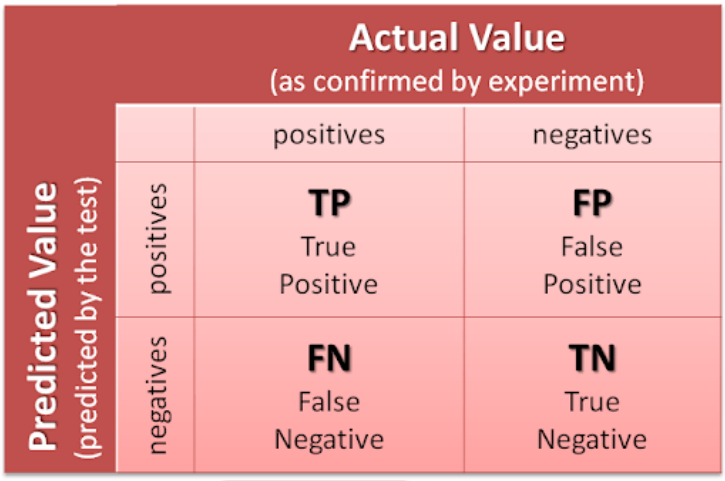
ตัววัดประสิทธิภาพของตัวแบบทำนายเป็นวิธีการที่ใช้ในการประเมินคุณภาพของการจัดกลุ่ม รวมไปถึงการประเมินประสิทธิภาพของตัวแบบทำนาย โดยการนำผลลัพธ์ที่ได้จากแบบจำลองมาสร้างเป็น Confusion Matrix ดังภาพประกอบที่ 9 ดังนี้

True Positive (TP) คือ สิ่งที่โปรแกรมทำนายว่า “จริง” และ มีค่าเป็น “ จริง ”

True Negative (TN) คือ สิ่งที่โปรแกรมทำนายว่า “ไม่จริง” และ มีค่า “ ไม่จริง ”

False Positive (FP) คือ สิ่งที่โปรแกรมทำนายว่า “จริง” แต่ มีค่าเป็น “ไม่จริง”

False Negative (FN) คือ สิ่งที่โปรแกรมทำนายว่า “ไม่จริง” แต่ มีค่าเป็น “จริง”



ภาพประกอบ 9 แสดงตาราง Confusion Matrix

ซึ่งการประเมินความแม่นตรงของตัวแบบทำนายที่เลือกใช้ประกอบด้วย

1) Accuracy คือ ค่าความถูกต้อง ค่าที่บอกว่าโปรแกรมสามารถทำนายได้แม่นยำขนาดไหนเป็นการวัดความถูกต้องของ Model โดยพิจารณารวมทุกคลาส

2) Precision คือ ค่าความแม่นยำ ค่าที่บอกว่าโปรแกรมทำนายว่าจริงถูกต้องเท่าไหร่ เป็นการวัดความแม่นยำของข้อมูล โดยพิจารณาแยกทีละคลาส

3) Recall คือ ค่าความระลึก ค่าที่บอกว่าโปรแกรมทำนายได้จริง เป็นอัตราส่วนเท่าไหร่ของค่าจริงทั้งหมดเป็นการวัดความถูกต้องของ Model โดยพิจารณาแยกทีละคลาส

4) f1_measure คือ ค่าความถ่วงดุล เป็นการนำค่า Precision กับ Recall มาคำนวณค่าเฉลี่ยกันสำหรับพิจารณาร่วมกันเพื่อความแม่นยำ และความครบถ้วน

2.9 โปรแกรม RapidMiner studio 9.5

RapidMiner Studio เป็นโปรแกรมที่ใช้สำหรับการออกแบบการวิเคราะห์ และประมวลผลข้อมูลผ่านทางหน้า GUI (Graphical User Interface) ซึ่งสามารถทำการจัดการข้อมูล และสร้างตัวแบบทำนายแบบต่าง ๆ ได้ง่าย และรวดเร็ว โดย Operator ที่ใช้ในการทำวิจัยประกอบด้วย Operator ดังนี้

2.9.1 Operator Read CSV ดังภาพประกอบที่ 10 เป็นตัวที่นำข้อมูลไปใช้งาน โดยถ้าข้อมูลไฟล์ csv มีการเปลี่ยนแปลงจะ update ไฟล์ให้อัตโนมัติ ไม่ต้องทำการโหลดใหม่ รายละเอียดดังนี้

ภาพประกอบ 10 Operator Read CSV

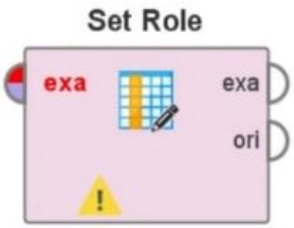
Input

- File (ไฟล์) เอกสารหรือไฟล์นามสกุล .CSV

Output

- Output (**ตารางข้อมูล**) พอร์ต (Port) นี้จะส่งออกชุดข้อมูล ExampleSet ที่สร้างจากไฟล์ CSV ที่นำเข้าจากพอร์ตอินพุต

2.9.2 Operator Set Role ดังภาพประกอบที่ 11 เป็นการกำหนด column left ให้เป็นประเภท Label รายละเอียดดังนี้



ภาพประกอบ 11 Operator Set Role

Input

- Example set (ตารางข้อมูล)

พอร์ตอินพุตจะรับค่าชุดข้อมูล ExampleSet

Output

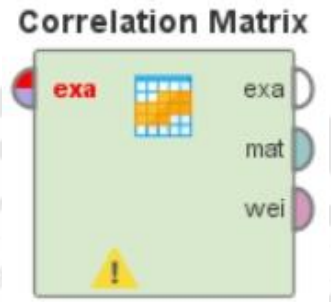
- Example set (ตารางข้อมูล)

พอร์ตเอาต์พุตจะส่งค่าชุดข้อมูล ExampleSet ที่กำหนดบทบาทเรียบร้อยแล้ว

- Original (ตารางข้อมูล)

พอร์ตนี้จะส่งค่าชุดข้อมูล ExampleSet ตั้งต้นที่ได้รับผ่านทางพอร์ตอินพุต โดยที่ไม่ได้ผ่านขั้นตอนการตั้งค่าบทบาทใด ๆ

2.9.3 Operator Correlation Matrix ดังภาพประกอบที่ 12 เป็นการนำข้อมูลมาหาความสัมพันธ์ระหว่างข้อมูล โดยสร้างค่าน้ำหนักให้ตัวแปร เพื่อหาว่าแต่ละตัวแปรมีความสัมพันธ์กันมากน้อยเพียงใด รายละเอียดดังนี้



ภาพประกอบ 12 Operator Correlation Matrix

Input

- example set (ตารางข้อมูล)

พอร์ตอินพุตจะรับค่าชุดข้อมูล ExampleSet

Output

- example set (ตารางข้อมูล)

พอร์ตนี้จะส่งค่าชุดข้อมูล ExampleSet ตั้งต้นที่ได้รับผ่านทางพอร์ตอินพุต โดยที่ไม่ได้ผ่านขั้นตอนการตั้งค่าบทบาทใด ๆ

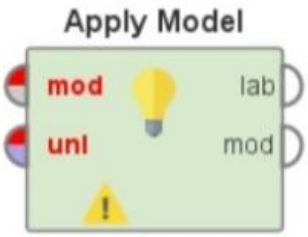
- matrix (ค่าเมทริกซ์)

พอร์ตนี้จะส่งค่าเมทริกซ์ของตัวแปร ซึ่งเมื่อไม่ได้กำหนดค่าของตัวแปรส่งผลให้ค่านั้นหายไป และพอร์ตจะคำนวณเฉพาะค่าที่สมบูรณ์

- weights (ค่าน้ำหนักของตัวแปร)

พอร์ตนี้จะส่งค่าน้ำหนักของตัวแปรที่คำนวณแล้วไปพอร์ตต่อไป

2.9.4 Operator Apply Model ดังภาพประกอบที่ 13 เป็นการนำ Model ที่ได้จากการประมวลผลข้อมูลในส่วนกลุ่มข้อมูลสอน (Training Data) มาใช้กับข้อมูลในกลุ่มข้อมูลทดสอบ (Test Data) ซึ่งแอตทริบิวต์ของข้อมูลที่ผ่านการเรียนรู้เพื่อสร้างแบบจำลองนั้นจะต้องมีค่าตรงกันกับชุดข้อมูล ExampleSet รายละเอียดดังนี้



ภาพประกอบ 13 Operator Apply Model

Input

- query set (ตารางข้อมูล)

พอร์ตอินพุตสำหรับรับค่า ExampleSet ที่มีค่าแอตทริบิวต์ของข้อมูลตรงกับแอตทริบิวต์ของข้อมูลที่ผ่านการเรียนรู้ของแบบจำลอง

- Model (แบบจำลอง)

พอร์ตอินพุตรับค่าแบบจำลอง ที่แอตทริบิวต์ของข้อมูลที่ผ่านการเรียนรู้เพื่อสร้างแบบจำลองนั้นมีค่าตรงกันกับชุดข้อมูล ExampleSet

Output

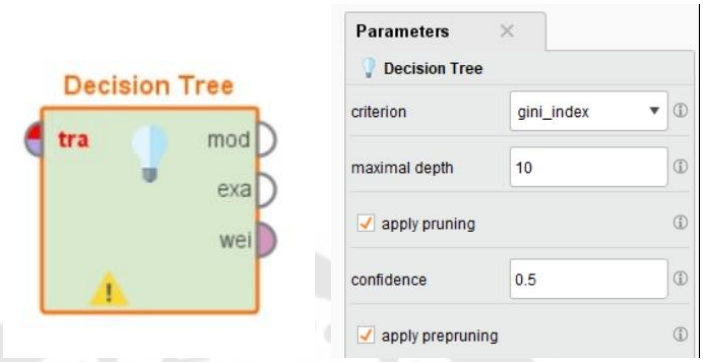
- result set (ตารางข้อมูล)

ExampleSet ที่ส่งมาจากพอร์ตนี้ถูกเปลี่ยนโดยใช้แบบจำลอง ในที่นี้จะเพิ่มค่ารายการแนะนำพร้อมการอันดับ

- Model (แบบจำลอง)

พอร์ตนี้จะส่งค่าแบบจำลองตั้งต้นที่ได้รับผ่านทางพอร์ตอินพุต โดยที่ไม่ได้ผ่านขั้นตอนการตั้งค่าบทบาทใด ๆ

2.9.5 Operator decision tree ดังภาพประกอบที่ 14 เป็นตัวที่ใช้สร้าง tree ตัวแบบทำนาย ที่ใช้ในการทำนายผลการเรียนในเทอมที่ 3 ของนักเรียนในข้อมูลตัวอย่าง โดยให้เชื่อมต่อกับไฟล์ Read CSV รายละเอียดดังนี้



ภาพประกอบ 14 Operator decision tree

Input

- example set (ตารางข้อมูล)

พอร์ตอินพุตจะรับค่าชุดข้อมูล ExampleSet

Output

- Model (แบบจำลอง)

พอร์ตเอาต์พุตให้ค่าแบบจำลอง decision tree

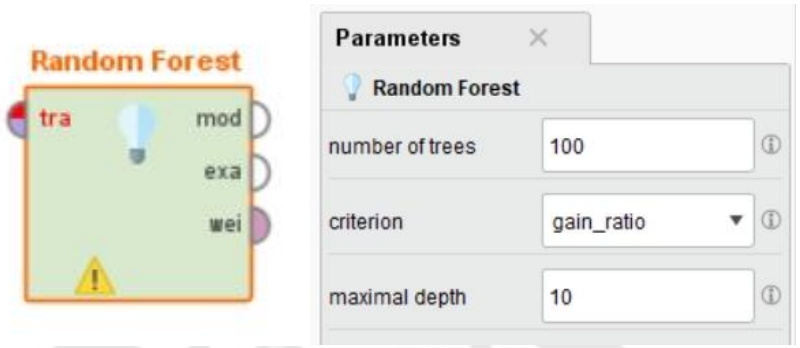
- example set (ตารางข้อมูล)

พอร์ตนี้จะส่งค่าชุดข้อมูล ExampleSet ตั้งต้นที่ได้รับผ่านทางพอร์ตอินพุต โดยที่ไม่ได้ผ่านขั้นตอนการตั้งค่าบทบาทใด ๆ

Parameters

- criterion การกำหนดเกณฑ์ที่จะเลือกแอตทริบิวต์เพื่อนำมาสร้างโมเดล
- maximal depth กำหนดค่าความลึกของต้นไม้แต่ละต้น จำนวน level ที่มากที่สุด ของ node ที่จะทำการ split
- confidence การกำหนดระดับความเชื่อมั่นที่ใช้สำหรับการคำนวณข้อผิดพลาดในสร้างโมเดล

2.9.6 Operator Random Forest ดังภาพประกอบที่ 15 เป็นตัวที่ใช้สร้างตัวแบบทำนายป่าไม้แห่งการทำนาย ที่ใช้ในการทำนายผลการเรียนในเทอมที่ 3 ของนักเรียนในข้อมูลตัวอย่าง โดยให้เชื่อมต่อกับไฟล์ Read CSV รายละเอียดดังนี้



ภาพประกอบ 15 Operator Random Forest

Input

- example Set (ตารางข้อมูล)

พอร์ตอินพุตจะรับค่าชุดข้อมูล ExampleSet ที่ใช้ในการเรียนรู้โมเดล

Output

- model (แบบจำลอง)

พอร์ตเอาต์พุตให้ค่าแบบจำลอง Naive Bayes

- example set (ตารางข้อมูล)

พอร์ตนี้จะส่งค่าชุดข้อมูล ExampleSet ตั้งต้นที่ได้รับผ่านทางพอร์ตอินพุต โดยที่ไม่ได้ผ่านขั้นตอนการตั้งค่าบทบาทใด ๆ

- weights (ค่าน้ำหนัก)

พอร์ตนี้จะส่งค่าชุดข้อมูล ExampleSet ที่กำหนดค่าน้ำหนักของแอตทริบิวต์ที่ใช้ในการทำนาย

Parameters

- number of tree การกำหนดจำนวน tree ที่จะนำมาสร้างโมเดล
- criterion การกำหนดเกณฑ์ที่จะเลือกแอตทริบิวต์เพื่อนำมาสร้างโมเดล
- maximal depth กำหนดค่าความลึกของต้นไม้แต่ละต้น จำนวน level ที่มากที่สุด ของ node ที่จะทำการ split

2.9.7 Operator Naive Bayes ดังภาพประกอบที่ 16 เป็นตัวที่ใช้สร้างตัวแบบทำนายความน่าจะเป็นที่ใช้ในการทำนายผลการเรียนในเทอมที่ 3 ของนักเรียนในข้อมูลตัวอย่าง โดยให้เชื่อมต่อกับไฟล์ Read CSV รายละเอียดดังนี้

ภาพประกอบ 16 Operator Naive Bayes

Input

- example set (ตารางข้อมูล)

พอร์ตอินพุตจะรับค่าชุดข้อมูล ExampleSet

Output

- Model (แบบจำลอง)

พอร์ตเอาต์พุตให้ค่าแบบจำลอง Naive Bayes

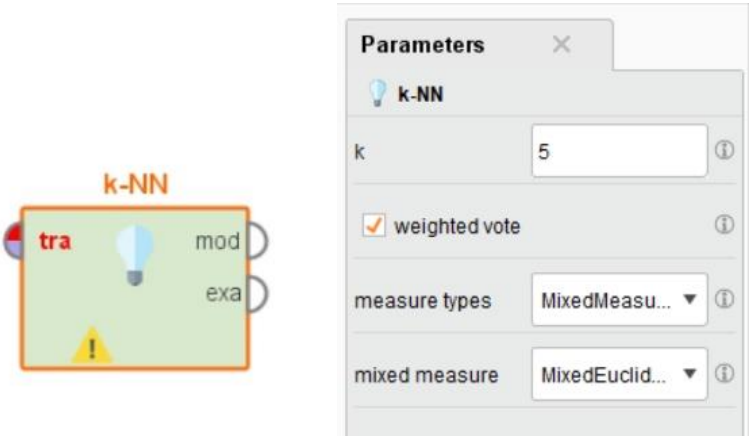
- example set (ตารางข้อมูล)

พอร์ตนี้จะส่งค่าชุดข้อมูล ExampleSet ตั้งต้นที่ได้รับผ่านทางพอร์ตอินพุต โดยที่ไม่ได้ผ่านขั้นตอนการตั้งค่าบทบาทใด ๆ

Parameters

- laplace correction คือการเพิ่มจำนวนเรคคอร์ด 1 เรคคอร์ดให้แต่ละกรณีของแต่ละค่าที่จะเกิดขึ้นในแอตทริบิวต์ที่พิจารณา เมื่อแอตทริบิวต์ที่พิจารณานั้นมีค่าไม่ครบจนทำให้การคำนวณความน่าจะเป็นในแต่ละครั้งมีเป็นค่า 0 เมื่อตั้งค่าแล้วจะเติมจำนวนเรคคอร์ดให้สามารถคำนวณค่าความน่าจะเป็นได้

2.9.8 Operator K-NN ดังภาพประกอบที่ 17 เป็นตัวที่ใช้สร้างตัวแบบทำนายโดยหาข้อมูลที่อยู่ใกล้เคียงเป็นจำนวน K ที่ใช้ในการทำนายผลการเรียนในเทอมที่ 3 ของนักเรียนในข้อมูลตัวอย่าง โดยให้เชื่อมต่อกับไฟล์ Read CSV รายละเอียดดังนี้



ภาพประกอบ 17 Operator k-NN

Input

- example set (ตารางข้อมูล)

พอร์ตอินพุตจะรับค่าชุดข้อมูล ExampleSet

Output

- Model (แบบจำลอง)

พอร์ตเอาต์พุตให้ค่าแบบจำลอง k-nn

- example set (ตารางข้อมูล)

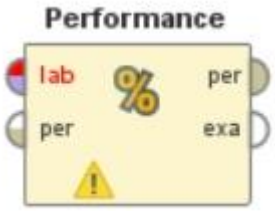
พอร์ตนี้จะส่งค่าชุดข้อมูล ExampleSet ตั้งต้นที่ได้รับผ่านทางพอร์ตอินพุต โดยที่ไม่ได้ผ่านขั้นตอนการตั้งค่าบทบาทใด ๆ

Parameters

- k การกำหนดจำนวนตัวอย่างที่อยู่ใกล้เคียงที่สุดกับตัวอย่างที่ไม่รู้จัก โดยทั่วไป k เป็นจำนวนเต็มบวก และเป็นจำนวนคี่

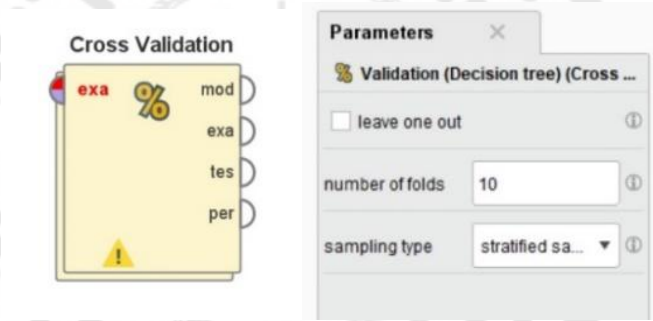
- measure types การกำหนดลักษณะการวัดที่ใช้ในการหาเพื่อนบ้านที่ใกล้ที่สุดและนำมาใช้ในการสร้างโมเดล

2.9.9 Operators Performance ดังภาพประกอบที่ 18 ใช้สำหรับวัดประสิทธิภาพของ Model ที่เลือกใช้ ซึ่งในที่นี้จะใช้ Decision Tree, Random forest, Naive Bayes, k-NN



ภาพประกอบ 18 Operator Set Role

2.9.10 Operators Cross Validation ดังภาพประกอบที่ 19 เป็นการตรวจสอบความถูกต้องของ Model ที่เลือกใช้ ซึ่งในที่นี้จะใช้ Decision Tree, Random forest, Naive Bayes, k-NN, ตรวจสอบด้วย 10-fold Cross Validation รายละเอียดดังนี้



ภาพประกอบ 19 Operators Cross Validation

Input

- Example set (ตารางข้อมูล)

พอร์ตอินพุตจะรับค่าชุดข้อมูล Example Set

Output

- Model (แบบจำลอง)

พอร์ตเอาต์พุตให้ค่าแบบจำลองตามอัลกอริทึมที่เชื่อมต่อ

Parameters

- number of folds จำนวนการแบ่งกลุ่มข้อมูลเพื่อใช้เป็นกลุ่มข้อมูลสอน (Training Data) และกลุ่มข้อมูลทดสอบ (Test Data)

- sampling type การกำหนดวิธีการเลือกกลุ่มตัวอย่างเพื่อทำการทำนาย

2.10 งานวิจัยที่เกี่ยวข้อง

Ferda Ünal ได้นำเสนองานวิจัยการเปรียบเทียบประสิทธิภาพของเทคนิคเหมืองข้อมูล ที่ใช้ในการทำนายผลการเรียนของเด็กนักเรียน โดยใช้ Dataset จาก UCI Dataset มีข้อมูล 33 คอลัมน์ ทำนายเกรดที่เด็กนักเรียนจะได้ ซึ่งจะแบ่งการให้เกรดเป็น 2 ประเภท คือ Five-level grading เป็นวิธีการแบ่งเกรดแบบให้เป็นช่วงคะแนน คือ คะแนน 16 - 20 ได้เกรด A, คะแนน 14 - 15 ได้เกรด B, คะแนน 12 - 13 ได้เกรด C, คะแนน 10 - 11 ได้เกรด D, คะแนน 0 - 9 ได้เกรด F และ Binary grading เป็นวิธีการแบ่งเกรดแบบ 2 ประเภท คือ คะแนน >,= 10 ได้ pass และคะแนนอื่น ๆ ที่ต่ำกว่า ได้ fail โดยใช้เทคนิคเหมืองข้อมูล 3 เทคนิค และยังใช้เทคนิคการคัดเลือกคุณลักษณะของ Feature ที่จะเลือกมาทำนายด้วยเทคนิค Wrapper พบว่าการให้เกรดแบบ Binary grading มีผลทำให้ประสิทธิภาพของการทำนายมีค่าสูงกว่าแบบ Five-level grading โดยเทคนิค Random forest เป็นเทคนิคที่ดีที่สุด และเมื่อใช้เทคนิค Wrapper พบว่า feature ที่มีความสำคัญในการทำนายมีทั้งหมด 13 feature ซึ่งคุณลักษณะทั้งหมดจะถูกจัดให้อยู่ในรูปของเซต หลังจากนั้นดำเนินการค้นหาเซตของคุณลักษณะที่เหมาะสมที่สุดก่อนเข้าสู่กระบวนการจำแนกข้อมูลโดยลำดับ (Sequential selection algorithm) ด้วยการพิจารณาคุณลักษณะที่เพิ่มทีละตัวตามลำดับ หรือลดคุณลักษณะลงทีละตัวตามลำดับจนกว่าจะได้เซตคุณลักษณะที่เหมาะสม

ชณิดาภา บุญประสม และ จรัญ แสนราช (2561) ได้นำเสนองานวิจัยเรื่อง การวิเคราะห์การทำนายการลาออกกลางคันของนักศึกษา ระดับปริญญาตรีโดยใช้เทคนิควิธีการทำเหมืองข้อมูล เพื่อวิเคราะห์หาปัจจัยที่เกี่ยวข้องในการลาออก สังเคราะห์ตัวแบบทำนายสำหรับการทำนาย และเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลของตัวแบบทำนายด้วยเทคนิคต้นไม้ตัดสินใจ (Decision Tree) การเรียนรู้เบย์ (Naïve-Bayes) และ K-NN (K-Nearest Neighbors) เมื่อวิเคราะห์ค่าน้ำหนักของแอททริบิวต์ด้วยวิธีการ Information Theory พบว่า มีปัจจัยที่เกี่ยวข้องในการลาออกกลางคันของนักศึกษาสูงสุด 5 ปัจจัย และนำมาทำการสร้างเป็นตัวแบบทำนายทดสอบผลลัพธ์ด้วยวิธีการ 10-Fold Cross Validation และวัดประสิทธิภาพด้วย ค่า Accuracy พบว่าตัวแบบทำนายที่สร้างด้วยเทคนิคการเรียนรู้เบย์ (Naïve-Bayes) มีประสิทธิภาพสูงสุด

ปริวรรต เพียรภายลุน, ธาดา จันตะคุณ, รักถิ่น เหลาหา (2560) ได้นำเสนองานวิจัยเรื่อง การเปรียบเทียบตัวแบบสำหรับใช้พยากรณ์คะแนน O-NET ด้วยเทคนิคเหมืองข้อมูล ใช้ข้อมูลของใช้ข้อมูลคะแนน O-NET ของนักเรียน ชั้นประถมศึกษาปีที่ 6 ที่ผ่านการทดสอบแล้ว จำนวน 148 คน ทำการคัดเลือกแอตทริบิวต์ที่เกี่ยวข้องมีทั้งหมด 7 แอตทริบิวต์ พบว่า ค่าความแม่นยำของตัวแบบ Naïve Bayes นั้นมีค่าความแม่นยำที่ดีที่สุด

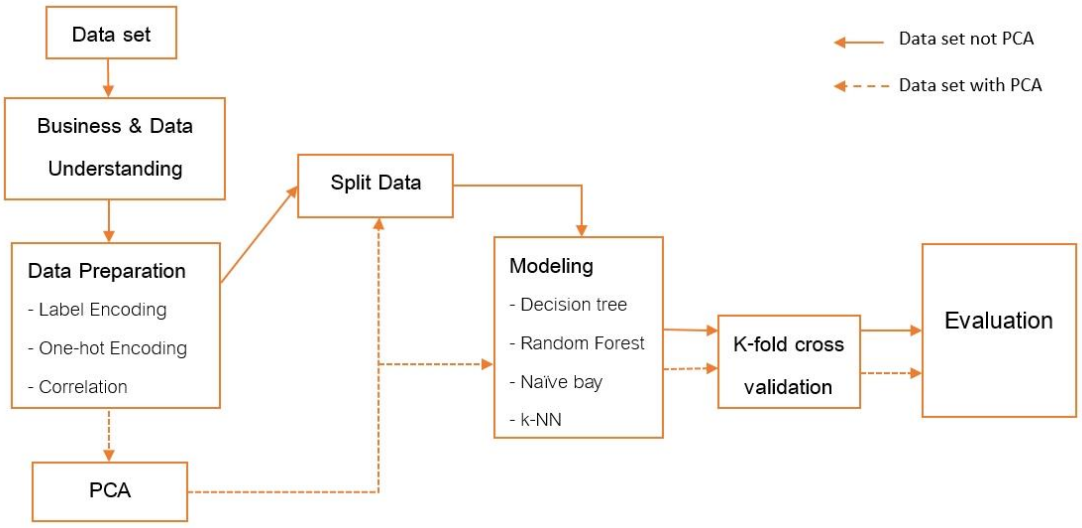
เสกสรรค์ วิลัยลักษณ์, วิภา เจริญภัณฑารักษ์, ดวงดาว วิชาดากุล (2558) ได้นำเสนองานวิจัยเรื่องการพัฒนาคลังข้อมูล และพยากรณ์ผลการเรียนของนักเรียน โดยใช้ข้อมูลนักเรียนระดับมัธยมศึกษาปีที่ 4 จำนวน 525 ระเบียน ประกอบด้วย16 คุณลักษณะ มาสร้างตัวแบบพยากรณ์ผลการเรียนโดยใช้เทคนิคเหมืองข้อมูลแบบโครงข่ายประสาทเทียมแบบมัลติเลเยอร์เพอร์เซ็ปตรอน (MLP) ซัพพอร์ตเวกเตอร์แมชชีน (SVM) และต้นไม้ตัดสินใจ (Decision Tree) มาสร้างตัวแบบพยากรณ์และทดสอบประสิทธิภาพของตัวแบบพยากรณ์ด้วย 10-fold Cross Validation ซึ่งพบว่าจากการคัดเลือกคุณลักษณะมี 5 ปัจจัยที่ส่งผลต่อผลการเรียน และชุดข้อมูลแบบไม่จัดกลุ่มมัลติเลเยอร์เพอร์เซ็ปตรอน ให้ค่าความถูกต้องสูงที่สุด

ศศิธร ตุ้มเพชรรัตน์, สมบูรณ์ อเนกฤทธิ์มงคล, สุมนา เกษมสวัสดิ์ (2560) ได้นำเสนองานวิจัยการสร้างตัวแบบพยากรณ์ผลการสอบ TOEIC ของนักศึกษาคณะศิลปศาสตร์ ที่สามารถสอบได้ 500 คะแนนขึ้นไป โดยใช้ซอฟต์แวร์ RapidMiner Studio 7.3 เป็นเครื่องมือในการวิเคราะห์ข้อมูลและสร้างตัวแบบทำนายการพยากรณ์ผลการสอบ ซึ่งพบว่ามีบางวิชาที่ไม่ได้เป็นปัจจัยในการมาสร้างตัวแบบทำนายใหม่ จึงปรับบางรายวิชาออก และเมื่อทดสอบตัวแบบพยากรณ์ด้วยวิธีการ10 Fold Cross Validation พบว่าตัวแบบที่สร้างด้วยวิธี K-NN (K-Nearest Neighbors) ให้ประสิทธิภาพสูงสุด

อัจฉราภรณ์ จุฑาผาด (2559) ได้นำเสนองานวิจัยเรื่อง การพัฒนาระบบสารสนเทศเพื่อการพยากรณ์จำนวนนักศึกษาใหม่ โดยใช้กฎการจำแนกต้นไม้ พยากรณ์จำนวนนักศึกษาใหม่ที่จะศึกษาในระดับป.ตรี ในปี พ.ศ. 2558 และ 2559 ซึ่งใช้ปัจจัยสำคัญที่นำมาสร้างตัวต้นแบบเพื่อการพยากรณ์ 5 ปัจจัย และผลการสอบคัดเลือก แบ่งเป็นคลาส คือ ผ่าน และไม่ผ่าน พบว่ามีการวัดค่าความถูกต้องได้เท่ากับร้อยละ 97.34 ค่าความแม่นยำเท่ากับร้อยละ 98.56 ค่าความระลึกเท่ากับร้อยละ 97.00 และค่าความถ่วงดุลเท่ากับร้อยละ 97.13

บทที่ 3
วิธีดำเนินการวิจัย

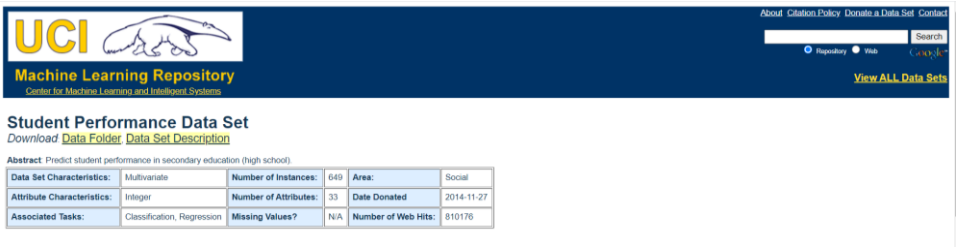
ในการดำเนินงานวิจัยครั้งนี้ได้เสนอวิธีการทำเหมืองข้อมูลโดยใช้โปรแกรม RapidMiner studio 9.5 ซึ่งมีกระบวนการตามมาตรฐานการทำเหมืองข้อมูลแบบ CRISP-DM ขั้นตอนการดำเนินงานวิจัยประกอบด้วย การทำความเข้าใจกับข้อมูลเพื่อให้รู้จักโครงสร้าง และรูปแบบของข้อมูล เมื่อเข้าใจข้อมูและรู้รูปแบบ หรือปัญหาของข้อมูลแล้ว จะนำข้อมูลไปแปลงให้อยู่ในรูปแบบที่สามารถนำไปใช้สร้างตัวแบบทำนายได้ พร้อมทั้งคัดเลือกเฉพาะข้อมูลที่สำคัญ และนำข้อมูลเข้าสู่โมเดล โดยจะมีการแบ่งข้อมูลออกเป็นข้อมูลที่ใช้ในการสอน และข้อมูลที่ใช้ทดสอบ เพื่อตรวจสอบประสิทธิภาพของโมเดล และประเมินผลโมเดล ดังภาพประกอบที่ 20



ภาพประกอบ 20 แสดงขั้นตอนการดำเนินงานวิจัย

3.1 การจัดเตรียมข้อมูล

เป็นการกลั่นกรองข้อมูลให้เหลือเพียงข้อมูลที่สมบูรณ์ และอยู่ในรูปแบบที่พร้อมจะนำไปประมวลผล โดยนำข้อมูลที่ได้จากแหล่งข้อมูล UCI Machine Learning Rapository ดังภาพประกอบที่ 21



ภาพประกอบ 21 รายละเอียดชุดข้อมูลของนักเรียน

ขั้นตอนในการจัดเตรียมข้อมูล รายละเอียดดังนี้

3.1.1 ทำความเข้าใจข้อมูลว่าในแต่ละแอตทริบิวต์มีความหมายอย่างไร ซึ่งข้อมูลที่ใช้ในงานวิจัยในครั้งนี้เป็นข้อมูลของนักเรียนระดับมัธยมศึกษาในโรงเรียนโปรตุเกสจำนวน 649 รายการ 33 ตัวแปร ซึ่งมีรายละเอียดดังตารางที่ 5 – 7 (Data Dictionary) และสามารถสำรวจข้อมูลโดยการ Import ข้อมูลเข้าใน Rapid Miner เพื่อให้ทราบโครงสร้างของข้อมูลเพื่อนำไปสู่การแปลงข้อมูลในขั้นตอนต่อไป โดยก่อนเข้าสู่กระบวนการแปลงข้อมูล ได้ทำการตัดข้อมูลออก 2 ตัวแปร ได้แก่ G1 และ G2 เนื่องจากเป็นเกรดของเทอมก่อนหน้าซึ่งจะไม่นำมาใช้ในการสร้างตัวแบบทำนาย

ตาราง 5 Data Dictionary อธิบายรายละเอียดในแต่ละแอตทริบิวต์

Attribute	Description	Type	Values
School	โรงเรียนที่นักเรียนเรียน	Binary	GP คือ Gabriel Pereira หรือ MS คือ Mousinho da Silveira)
Sex	เพศของนักเรียน	Binary	F คือ เพศหญิง(female) หรือ M คือ เพศชาย(male)
Age	อายุของนักเรียน	Numeric	ตั้งแต่ 15 ปี ถึง 22 ปี
Address	ที่อยู่ของนักเรียน	Binary	U คือ ในเมือง(Urban) หรือ R คือ ชนบท(Rural)
Famsize	จำนวนสมาชิกในครอบครัว	Binary	LE3 คือ น้อยกว่า หรือ เท่ากับ 3 และ GT3 คือ มากกว่า 3
Pstatus	สถานะของครอบครัว	Binary	T คือ อยู่ด้วยกัน หรือ A คือ แยกกันอยู่
Medu	ระดับการศึกษาของแม่	Numeric	0 คือ ไม่ได้เรียน, 1 คือ เรียนถึงเกรด 4, 2 คือ เรียนถึงเกรด 5 -9, 3 ระดับมัธยมศึกษา หรือ 4 ระดับที่สูงกว่า
Fedu	ระดับการศึกษาของพ่อ	Numeric	0 คือ ไม่ได้เรียน, 1 คือ เรียนถึงเกรด 4, 2 คือ เรียนถึงเกรด 5 -9, 3 ระดับมัธยมศึกษา หรือ 4 ระดับที่สูงกว่า
Mjob	อาชีพของแม่	Nominal	teacher, health care related, civil services, at_home, other

ตาราง 6 (ต่อ)

Attribute	Description	Type	Values
Fjob	อาชีพของพ่อ	Nominal	teacher, health care related, civil services, at_home, other
Guardian	ผู้ปกครองของนักเรียน	Nominal	mother คือ แม่, father คือ พ่อ, other คือ อื่น ๆ
Traveltime	ระยะเวลาในการเดินทางไปโรงเรียน	Numeric	1 คือ น้อยกว่า 15 นาที, 2 คือ 15 ถึง 30 นาที, 3 คือ 30 นาที – 1ชั่วโมง หรือ 4 คือ มากกว่า 1 ชั่วโมง
Studytime	ระยะเวลาในการเรียนต่อ 1 สัปดาห์	Numeric	1 คือ น้อยกว่า 2 ชั่วโมง, 2 คือ 2 ถึง 5 ชั่วโมง, 3 คือ 5 – 10ชั่วโมง หรือ 4 คือ มากกว่า 10 ชั่วโมง
Failures	ซ้ำชั้น/ไม่ผ่าน กี่ครั้ง	Numeric	ตั้งแต่ 1 - 4 ครั้ง
Schoolsup	เรียนพิเศษที่ รร.	Binary	yes หรือ no
Famsup	เรียนพิเศษที่บ้าน	Binary	yes หรือ no
Paid	จ่ายเงินเพื่อเรียนพิเศษ	Binary	yes หรือ no
Activities	กิจกรรมนอกหลักสูตร	Binary	yes หรือ no
Nursery	เรียนอนุบาล	Binary	yes หรือ no
Higher	ต้องการเรียนสูงๆ	Binary	yes หรือ no
Internet	ที่บ้านมีอินเตอร์เน็ต	Binary	yes หรือ no
Romantic	อยู่ในความสัมพันธ์ที่โรแมนติก/มีแฟน	Binary	yes หรือ no
Famrel	ระดับความผูกพันธ์ในครอบครัว	Numeric	ระดับ 1 คือ แย่มาก ถึง 5 คือ ดีเยี่ยม
Freetime	มีเวลาว่างหลังเลิกเรียนแค่ไหน	Numeric	ระดับ 1 คือ น้อยมาก ถึง 5 คือ สูงมาก
Gout	ออกไปเที่ยวกับเพื่อนบ่อยแค่ไหน	Numeric	ระดับ 1 คือ น้อยมาก ถึง 5 คือ สูงมาก
Dalc	การดื่มแอลกอฮอล์ในวันจันทร์-ศุกร์	Numeric	ระดับ 1 คือ น้อยมาก ถึง 5 คือ สูงมาก
Walc	การดื่มแอลกอฮอล์ในหยุด	Numeric	ระดับ 1 คือ น้อยมาก ถึง 5 คือ สูงมาก
Health	ระดับของสุขภาพ	Numeric	ระดับ 1 คือ แย่มาก ถึง 5 คือ ดีเยี่ยม

ตาราง 7 (ต่อ)

Attribute	Description	Type	Values
absences	ขาดเรียนกี่ครั้ง	Numeric	ตั้งแต่ 0 – 93 ครั้ง
G1	เกรดเทอมที่ 1	Numeric	ตั้งแต่ 0 ถึง 20 คะแนน
G2	เกรดเทอมที่ 2	Numeric	ตั้งแต่ 0 ถึง 20 คะแนน
G3	เกรดเทอมที่ 3	Numeric	ตั้งแต่ 0 ถึง 20 คะแนน

3.1.2 เนื่องจากข้อมูลที่นำมาใช้ มีทั้งในรูปแบบของตัวอักษร(Nominal) และตัวเลข (Numeric) จึงทำการแปลงข้อมูลจากตัวอักษรให้อยู่ในรูปแบบของตัวเลขทั้งหมดเพื่อให้ตัวแบบทำนายสามารถทำงานได้ ดังนี้

1 . Label endcoding: คือการแปลงข้อมูลในแต่ละคอลัมน์ให้เป็น id ที่ต่างกัน ซึ่งตัวเลขที่อยู่ใกล้กันจะถูกตีความว่ามี "คุณค่า" ใกล้กัน เช่น 1 กับ 2 ใกล้กันมากกว่า 1 กับ 3 ดังนั้น ถ้าค่าจริงของ 1, 2, 3 มีความสัมพันธ์เชิงคุณค่าที่ต่อเนื่องกัน สามารถใช้ได้เลย เช่นถ้า 1 แปลว่า "น้อย", 2 แปลว่า "ปานกลาง", และ 3 แปลว่า "มาก" เป็นต้น เช่น Famsize คือ จำนวนสมาชิกในครอบครัว มีค่าเป็น LE3 คือ น้อยกว่า หรือ เท่ากับ 3 และ GT3 คือ มากกว่า 3 แปลงให้อยู่ในรูป Label endcoding ดังตารางที่ 8

ตาราง 8 ตัวอย่างการแปลงข้อมูลในรูป Label encoding

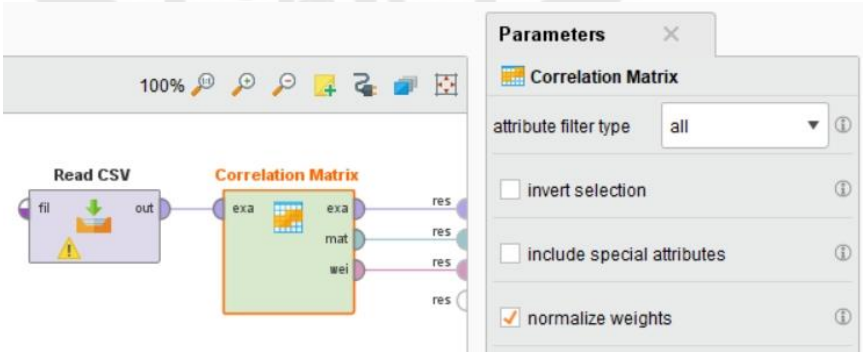
Famsize	Label encoding
LE3	1
GT3	2

2. One Hot Endcoding : เป็นการแปลงข้อมูลประเภทข้อความให้แตกเป็นคอลัมน์ย่อย ๆ แบบ Binary 0/1 กรณีที่ค่าจริงของแต่ละเลขนั้นไม่ได้มีความสัมพันธ์ต่อเนื่องเป็นลำดับขั้นกัน เช่น Mjob คือ อาชีพของแม่ ได้แก่ teacher, health care related, civil services, at_home, other ค่าของแต่ละเลขล้วนมีความหมายของตัวเอง และไม่เกี่ยวข้องทางลำดับชั้นกับค่าอื่น แปลงให้อยู่ในรูป one hot endcoding ดังตารางที่ 9

ตาราง 9 ตัวอย่างการแปลงข้อมูลในรูป one hot encoding

teacher	health care related	civil service	at_home	other
1	0	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
0	0	1	0	1

3.1.3 การหาความสัมพันธ์ของตัวแปรก่อนการทำนาย (Correlation) เป็นการหาความสัมพันธ์ของแต่ละตัวแปร เพื่อเลือกนำเฉพาะตัวแปรที่จะส่งผลให้การทำนายมีความถูกต้องแม่นยำมากยิ่งขึ้น โดยเริ่มจากการ Import ข้อมูลเข้าใน Rapid Miner โดยใช้ Operator Correlation Matrix และกำหนด Parameters attribute filter type .ให้เป็น all เพื่อเลือกข้อมูลทั้งหมดในการทำ Correlation และเลือก normalize weights เพื่อประมวลค่าน้ำหนัก Correlation ให้เป็นค่าตั้งแต่ 0 – 1 และ 0 – (-1) ดังภาพประกอบที่ 22



ภาพประกอบ 22 การตั้งค่า Parameter ภายใน Operator Correlation Matrix

3.2 การสร้างตัวแบบทำนายใน Rapid Miner studio 9.5

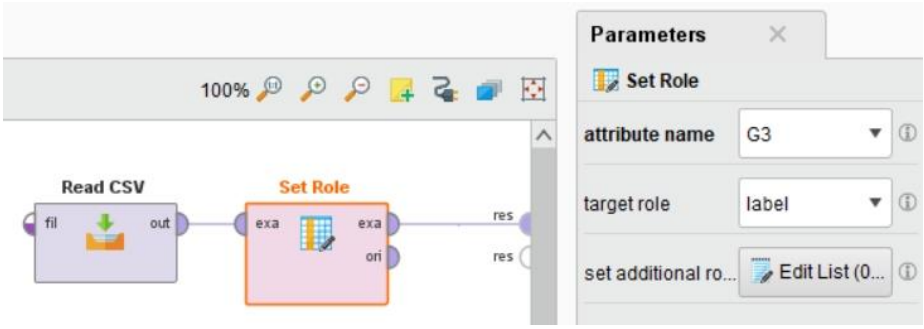
เป็นการสร้างและทดสอบความถูกต้องของตัวแบบทำนายโดยการวิเคราะห์ข้อมูล ประมวลผลตัวแบบทำนายจากข้อมูลทั้งหมด เริ่มจากการ Import ข้อมูล เข้าใน Rapid Miner studio 9.5 ดังภาพประกอบที่ 23 เพื่อวิเคราะห์และคำนวณค่าต่าง ๆ ที่ใช้ในการทำนาย

	ırel ger	freetime integer	goout integer	Dalc integer	Walc integer	health integer	absences integer	G3 polynominal label
1		3	4	1	1	3	4	pass
2		4	4	1	2	2	6	pass
3		2	2	1	3	5	6	pass
4		1	3	1	3	4	0	pass
5		3	3	1	1	4	10	pass
6		2	3	1	2	1	0	pass
7		4	2	1	1	5	0	pass
8		2	2	3	3	5	14	fail
9		3	5	2	4	4	0	pass
10		3	4	1	1	4	0	pass
11		2	5	1	2	5	0	pass

ภาพประกอบ 23 การกำหนดชนิดของข้อมูลที่ต้องการทำนาย

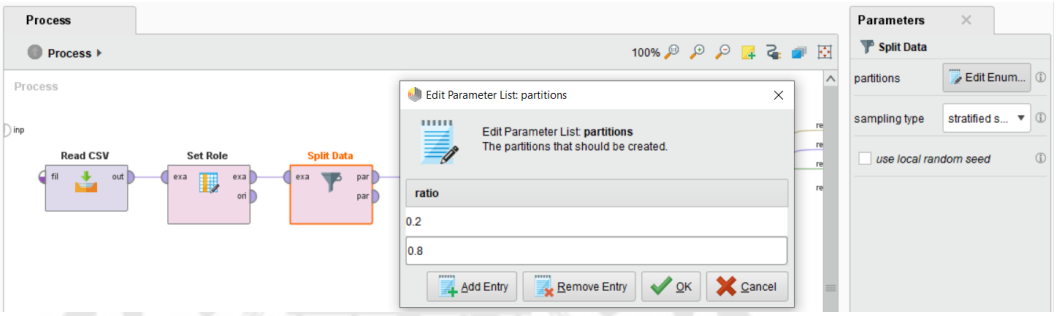
จากการสำรวจข้อมูลประกอบด้วย 649 Row 28 ตัวแปร ซึ่งแต่ละ ตัวแปร จะแสดงให้เห็นถึงคุณลักษณะเฉพาะของข้อมูล คือ integer จากภาพประกอบที่ 18 เป็นการแสดงค่าสถิติของข้อมูลในแต่ละ ตัวแปร ซึ่งไม่พบข้อมูลว่าง หรือข้อมูลที่ผิดแปลก และสามารถนำมาสร้างตัวแบบทำนายในการทำนายได้

3.2.1 การสร้างตัวแบบทำนายเพื่อทำนายผลการเรียนในเทอมที่ 3 ของนักเรียน โดยนำเข้าชุดข้อมูล student_por.csv ผ่านโอเปอร์เรเตอร์ ReadCSV เรียบร้อยแล้ว จึงกำหนดบทบาทของแอตทริบิวต์ โดยใช้โอเปอร์เรเตอร์ Set Role ระบุให้ ตัวแปร G3 เป็น Label เพราะเป็นข้อมูลที่ต้องการทำนาย ดังภาพประกอบที่ 24



ภาพประกอบ 24 การตั้งค่า Parameter ภายใน Operator Set Role

จากนั้นแบ่งข้อมูลเป็น 2 กลุ่ม ผ่านโอเปอร์เรเตอร์ Split Data ดังภาพประกอบที่ 25 คือ กลุ่มข้อมูลที่จะใช้สอน (Training Data) 80% และข้อมูลที่จะใช้ทดสอบ (Test Data) 20% และกำหนด Parameters ให้แบ่งข้อมูลแบบ Stratified sampling เพื่อสุ่มตัวอย่างโดยแยกประชากรออกเป็นกลุ่มประชากรย่อย ๆ ก่อน เพื่อให้ได้จำนวนกลุ่มตัวอย่างตามสัดส่วนของขนาดกลุ่มตัวอย่างและกลุ่มประชากร ก่อนการเชื่อมต่อกับเทคนิคการทำนายต่าง ๆโดยใช้เทคนิคเหมืองข้อมูล ดังนี้



ภาพประกอบ 25 การ Split ข้อมูลเป็น Training Data และ Test data

1. การสร้างตัวแบบทำนายโดยใช้เทคนิคต้นไม้ตัดสินใจ (Decision Tree) โดยใช้ Operator Decision Tree และกำหนด Parameters ดังภาพประกอบที่ 26

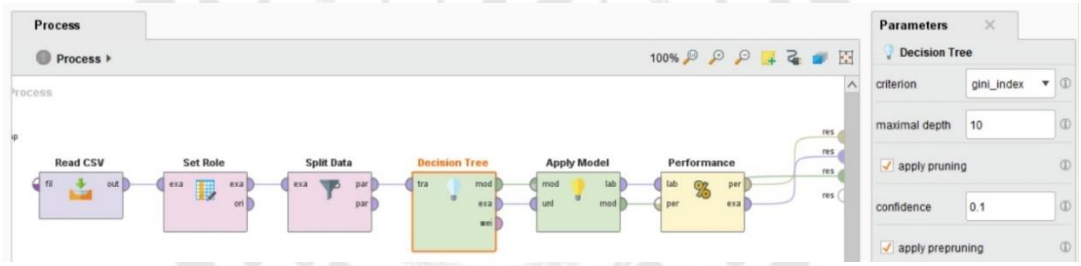
- criterion ตั้งค่าเป็น gini_index ค่าที่บ่งบอกว่าคุณลักษณะหรือปัจจัยใดควรนำมาใช้เป็นคุณลักษณะในการแบ่งกลุ่มของอัลกอริธึม โดยจะคำนวณให้ได้ค่า gini ที่น้อยที่สุดมาเป็นโหนดราก

- confidence ตั้งค่าเป็น 0.1 เพื่อกำหนดค่าช่วงความเชื่อมั่นที่ใช้สำหรับการคำนวณ error ในการทำนาย

- maximal depth ตั้งค่าเป็น 10 ขั้น เพื่อกำหนดค่าความลึกของต้นไม้ โดยได้จากการเปรียบเทียบการกำหนดค่า maximal depth ที่ให้ผลลัพธ์ที่ดีที่สุด เมื่อถึง Scenario 5 ที่ให้ผลลัพธ์เท่ากับค่าใน Scenario 2- 4 ผู้วิจัยจึงหยุดสุ่มค่า maximal depth ดังตารางที่ 10

ตาราง 10 แสดงผลลัพธ์จากการทดสอบสุ่มหาค่า maximal depth ที่ดีที่สุดของเทคนิคต้นไม้ตัดสินใจ (Decision Tree)

No.	maximal depth	ค่าความถูกต้อง %
Scenario 1	5	90.00
Scenario 2	10	91.54
Scenario 3	15	91.54
Scenario 4	20	91.54
Scenario 5	25	91.54



ภาพประกอบ 26 การสร้างตัวแบบทำนายและการตั้งค่า Parameter Decision Tree

2. การสร้างตัวแบบทำนาย โดยเทคนิคป่าแห่งการทำนาย (Random Forest) โดยใช้ Operator RandomForest และกำหนด Parameters ดังภาพประกอบที่ 27

- criterion ตั้งค่าเป็น gini_index ค่าที่บ่งบอกว่าคุณลักษณะหรือปัจจัยใดควรนำมาใช้เป็นคุณลักษณะในการแบ่งกลุ่มของอัลกอริธึม โดยจะคำนวณให้ได้ค่า gini ที่น้อยที่สุดมาเป็นโหนดราก

- number of trees เพื่อกำหนดจำนวนต้นไม้ หรือโมเดลที่นำมาสร้างให้เป็นป่าแห่งทำนาย ตั้งค่าเป็น 100 ต้น โดยได้จากการเปรียบเทียบการกำหนดค่า number of trees ที่ให้ผลลัพธ์ที่ดีที่สุด ซึ่งเมื่อถึง Scenario 4 ให้ผลลัพธ์เท่ากับค่าใน Scenario 2- 3 ผู้วิจัยจึงหยุดสุ่มค่า number of trees ดังตารางที่ 11

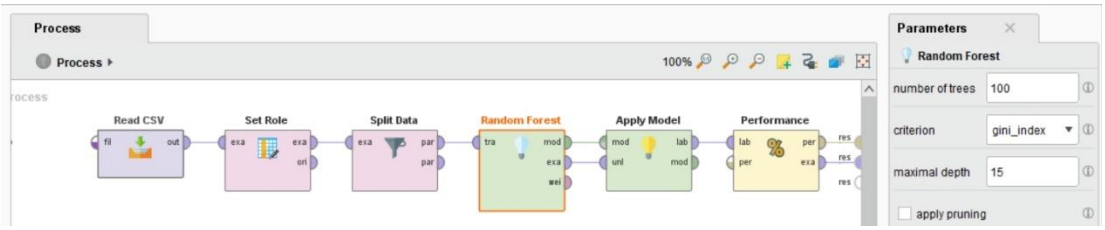
ตาราง 11 แสดงผลลัพธ์จากการทดสอบสุ่มหาค่า number of trees ที่ดีที่สุดของป่าแห่งการทำนาย (Random Forest)

No.	number of trees	ค่าความถูกต้อง %
Scenario 1	50	99.23
Scenario 2	100	100
Scenario 3	150	100
Scenario 4	200	100

- maximal depth เพื่อกำหนดค่าความลึกของต้นไม้แต่ละต้น จำนวน level ที่มากที่สุด ของ node ที่จะทำการ split ตั้งค่าเป็น 15 ขั้น โดยได้จากการเปรียบเทียบการกำหนดค่า maximal depth ที่ให้ผลลัพธ์ที่ดีที่สุด เมื่อถึง Scenario 4 ที่ให้ผลลัพธ์เท่ากับค่าใน Scenario 2 - 3 ผู้วิจัยจึงหยุดสุ่มค่า maximal depth ดังตารางที่ 12

ตาราง 12 แสดงผลลัพธ์จากการทดสอบสุ่มหาค่า maximal depth ที่ดีที่สุดของป่าแห่งการทำนาย (Random Forest)

No.	maximal depth	ค่าความถูกต้อง %
Scenario 1	5	88.46
Scenario 2	10	98.46
Scenario 3	15	100
Scenario 4	20	100



ภาพประกอบ 27 การสร้างตัวแบบทำนายและการตั้งค่า Parameter Random Forest

3. การสร้างตัวแบบทำนาย โดยใช้เทคนิคการเรียนรู้เบย์ (Naïve-Bayes)

โดยใช้ Operator RandomForest และกำหนด Parameters ดังภาพประกอบที่ 28

- laplace correction การตั้งค่าให้การทำนายความน่าจะเป็นในแต่ละครั้งไม่เป็นค่า 0

ภาพประกอบ 28 การสร้างตัวแบบทำนายและการตั้งค่า Parameter Naïve-Bayes

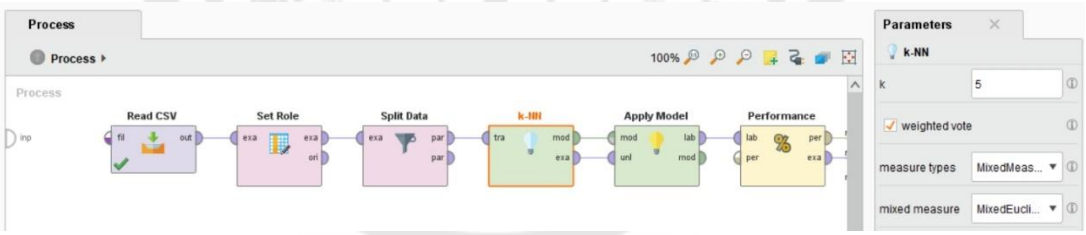
4. การสร้างตัวแบบทำนายโดยใช้เทคนิคเพื่อนบ้านใกล้ที่สุด k-NN (K-Nearest Neighbors) โดยใช้ Operator k-NN และกำหนด Parameters ดังภาพประกอบที่ 29

- k เป็นการกำหนดข้อมูลที่อยู่ใกล้เคียง ตั้งค่าเป็น 5 โดยได้จากการเปรียบเทียบการกำหนดค่า k ที่ให้ผลลัพธ์ที่ดีที่สุด เมื่อถึง Scenario 5 ที่ให้ผลลัพธ์เท่ากับค่าใน Scenario 3 - 4 ผู้วิจัยจึงหยุดสุ่มค่า maximal depth ดังตารางที่ 13

ตาราง 13 แสดงผลลัพธ์จากการทดสอบสุ่มหาค่า k ที่ดีที่สุดของป่าแห่งการทำนาย เทคนิคเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbors)

No.	k	ค่าความถูกต้อง %
Scenario 1	5	82.31
Scenario 2	10	81.54
Scenario 3	15	77.69
Scenario 4	20	77.69
Scenario 5	25	77.69

- measure types ตั้งค่าเป็น mixed measures หมายถึงให้สามารถคำนวณระยะทางในกรณีที่แอตทริบิวต์นั้นเป็นค่าที่ต่างกัน ทั้งที่เป็นตัวเลขและที่ไม่ใช่ตัวเลข ให้สามารถหาระยะทางได้



ภาพประกอบ 29 การสร้างตัวแบบทำนายการ และการตั้งค่า Parameter k-NN

3.2 การคัดเลือกข้อมูล

เป็นการนำข้อมูลมาพิจารณาว่าจะเลือกข้อมูลใดบ้าง เนื่องจากการสร้างตัวแบบทำนายเป็นการนำข้อมูลทั้งหมดมาใช้ในการทำนาย โดยหลังจากการผ่านขั้นตอนการเตรียมข้อมูล และหาความสัมพันธ์ของตัวแปรแล้ว ทำให้มีตัวแปร ทั้งหมด 42 ตัวแปร ซึ่งยังไม่ทราบว่าข้อมูลที่ใช้ จะสามารถใช้งานได้ทั้งหมดหรือไม่ อาจส่งผลต่อความถูกต้องของการประมวลผล ดังนั้นผู้วิจัยจึงทำการคัดเลือกข้อมูล โดยการวิเคราะห์องค์ประกอบหลัก Principal Component Analysis (PCA) ซึ่งเป็นเทคนิคเชิงคำนวณที่นิยมนำมาใช้ในการย่อยข้อมูล และวิเคราะห์ข้อมูลที่มีหลายตัวแปร

เพื่อหาความสัมพันธ์ของตัวแปรเหล่านั้น และปรับให้ข้อมูลมีความกระชับ มีขนาดเล็กลง เมื่อผ่านกระบวนการ PCA แล้ว ส่งผลให้มี ตัวแปร คงเหลือ 22 ตัวแปร ดังภาพประกอบที่ 30

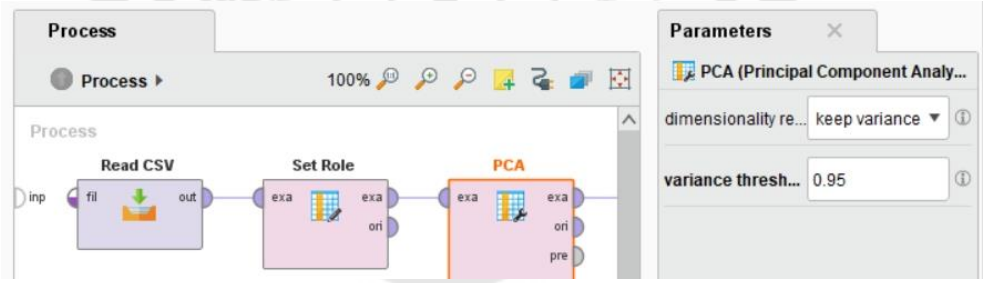
pc_9	pc_10	pc_11	pc_12	pc_13	pc_14	pc_15	pc_16	pc_17	pc_18	pc_19	pc_20	pc_21	pc_22
).400	-0.314	-1.388	0.612	-0.611	0.694	-0.195	0.409	0.148	0.245	-0.258	0.311	-0.631	-0.184
1.565	2.057	0.756	-0.582	1.112	-0.686	0.032	-0.366	-0.490	0.994	-0.165	0.250	-0.168	0.684

ภาพประกอบ 30 จำนวนตัวแปรคงหลือเมื่อผ่านกระบวนการ PCA

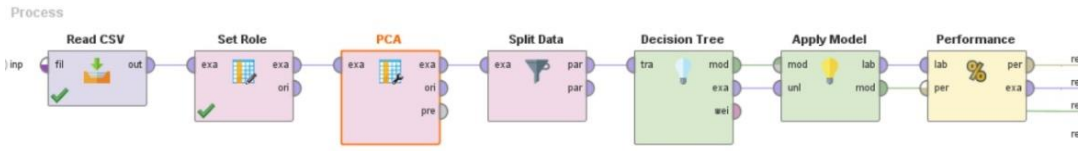
จากนั้นสร้างตัวแบบทำนายให้ครบทั้ง 4 เทคนิค ดังภาพประกอบที่ 32-35 โดยมีการตั้งค่า Parameter เช่นเดียวกับการสร้างตัวแบบทำนายซึ่งใช้ข้อมูลที่ยังไม่ได้ผ่านกระบวนการ PCA แต่เพิ่ม Operator PCA และกำหนด Parameters ดังภาพประกอบที่ 31

- dimensionality reduction ตั้งค่าเป็น keep variance โดยเป็นการกำหนดช่วงความแปรปรวน ที่ใช้วัดการกระจายของข้อมูลได้

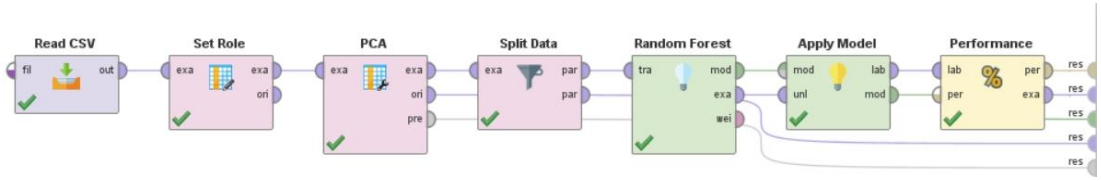
- variance threshold เป็นการกำหนดค่าที่ต้องการ โดยตั้งค่าเป็น 0.95 คือ ตั้งค่าความแปรปรวนของตัวแปรซึ่งจะเลือกเอาตัวแปรที่มีค่า variance ไม่เกินที่ 0.95



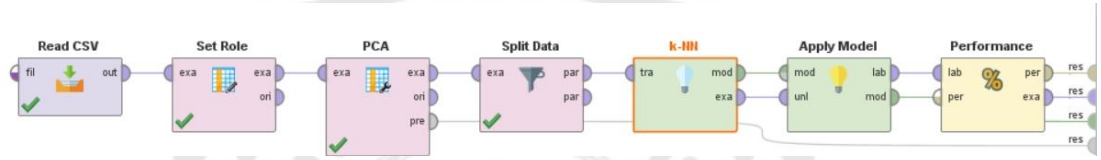
ภาพประกอบ 31 การตั้งค่า Parameter Operator PCA



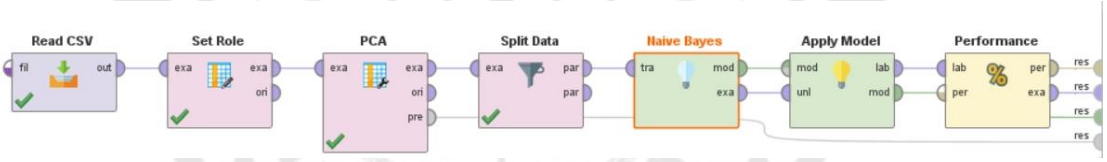
ภาพประกอบ 32 การสร้างตัวแบบทำนาย Decision Tree จากข้อมูลที่ผ่าน PCA



ภาพประกอบ 33การสร้างตัวแบบทำนาย Random Forest จากข้อมูลที่ผ่าน PCA



ภาพประกอบ 34 การสร้างตัวแบบทำนาย k-NN จากข้อมูลที่ผ่าน PCA



ภาพประกอบ 35 การสร้างตัวแบบทำนาย Naïve-Bayes จากข้อมูลที่ผ่าน PCA

3.3 การวัดประสิทธิภาพของตัวแบบทำนาย

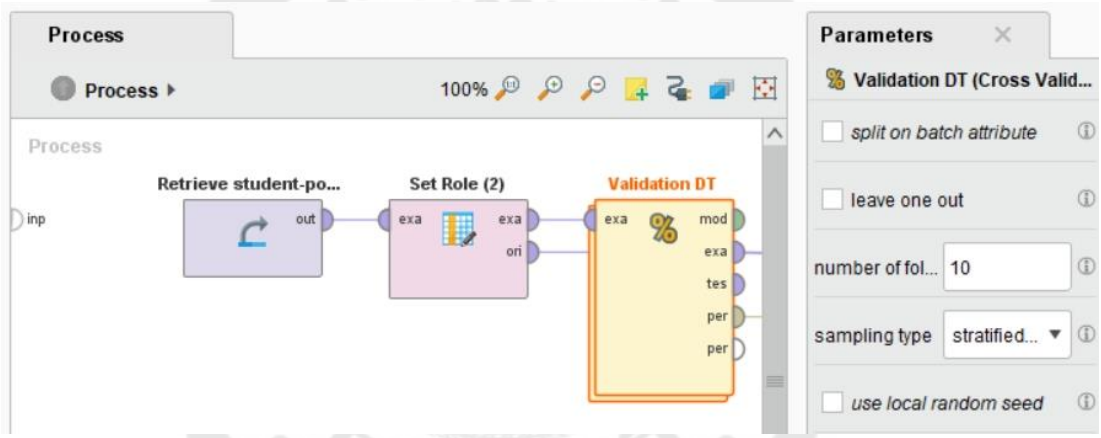
เป็นการทำนายค่าความแม่นยำของตัวแบบทำนายแต่ละประเภท โดยใช้วิธี K-fold cross validation ซึ่งจะทำการแบ่งข้อมูลเป็น 10 ส่วน โดยเท่า ๆ กัน แล้วทำการทดสอบ จากนั้นจะเปลี่ยนข้อมูลทดสอบชุดที่ 2 เป็นข้อมูลทดสอบ และข้อมูลส่วนที่เหลือคือ 1-(k-1) เป็นชุดข้อมูลสอน ทำไปเรื่อย ๆ จนครบถึงชุดทดสอบที่ K เป็นชุดทดสอบ และส่วนที่ 1-(k-1) เป็นชุดข้อมูลสอน และนำค่าที่ได้มาหาค่าเฉลี่ย

โดยสร้างตัวแบบทำนายเพื่อทำนายผลการเรียนในเทอมที่ 3 ของนักเรียน โดยใช้เทคนิคเหมืองข้อมูล เช่นเดิมอีกครั้ง และใช้ชุดข้อมูลทั้งที่ผ่านและไม่ผ่านกระบวนการ PCA มาทำการวัดค่าความแม่นยำของตัวแบบทำนายแต่ละประเภท โดยใช้วิธี K-fold cross validation

จากนั้นสร้างตัวแบบทำนายให้ครบทั้ง 4 เทคนิค ดังภาพประกอบที่ 28 โดยใช้ข้อมูลทั้งที่ผ่าน และไม่ผ่านกระบวนการ PCA รวมทั้งตั้งค่า Parameter เช่นเดิม แต่เพิ่ม Operator Validation และกำหนด Parameters ดังภาพประกอบที่ 36

- number of fold ตั้งค่าข้อมูลในการทดสอบเป็น 10 ส่วนเท่า ๆ กัน

- sampling type ตั้งค่าเป็น Stratified sampling เพื่อสุ่มตัวอย่างโดยแยกประชากรออกเป็นกลุ่มประชากรย่อย ๆ ก่อน การเลือกตัวอย่างวิธีนี้ ประชากรจะถูกแบ่งออกเป็นชั้นภูมิ (Stratum) ตามลักษณะอย่างใดอย่างหนึ่ง โดยไม่ให้มีหน่วยซ้ำกัน คือแต่ละหน่วยในประชากรจะต้องอยู่ในชั้นภูมิใดชั้นภูมิหนึ่งเท่านั้น โดยที่พยายามจัดให้ชั้นภูมิเดียวกันประกอบด้วยหน่วยที่มีลักษณะคล้ายคลึงกันมากที่สุด และมีความแตกต่างระหว่างชั้นภูมิมากที่สุด



ภาพประกอบ 36 การตั้งค่า Parameter Operator Validation

บทที่ 4
ผลการศึกษา

งานวิจัยนี้เป็นงานวิจัยเพื่อศึกษาการสร้างตัวแบบทำนายการพยากรณ์ผลสัมฤทธิ์ทางการเรียน โดยใช้ซอฟต์แวร์ Rapid Miner Studio โดยผู้วิจัยได้ทำการวิเคราะห์ข้อมูล และนำเสนอผลการศึกษาดังนี้

1. ผลการเตรียมข้อมูล (Data Preparation) เพื่อให้ข้อมูลพร้อมในการสร้างตัวแบบทำนาย
2. ผลการศึกษาการสร้างตัวแบบทำนายการพยากรณ์โดยใช้เทคนิคต่าง ๆ ดังนี้
    - เทคนิคต้นไม้ตัดสินใจ (Decision Tree)
    - เทคนิคป่าแห่งการทำนาย (Random Forest)
    - เทคนิคการเรียนรู้เบย์ (Naïve-Bayes)
    - เทคนิคเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbors)

2. ผลการศึกษาการสร้างตัวแบบทำนายการพยากรณ์ด้วยเทคนิคการวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis : PCA)

3. ผลการวิเคราะห์ความแม่นตรง และเปรียบเทียบประสิทธิภาพของตัวแบบทำนายด้วยเทคนิคต่าง ๆ โดยใช้วิธี K-fold cross validation

4.1 ผลการเตรียมข้อมูล (Data Preparation) เพื่อให้ข้อมูลพร้อมในการสร้างตัวแบบทำนาย

ผู้วิจัยได้นำข้อมูลเข้าสู่ตัวแบบทำนายโดย ใช้ซอฟต์แวร์ Rapid Miner Studio ในการวิเคราะห์ข้อมูล ซึ่งข้อมูลที่ใช้นั้นผ่านขั้นตอนการเตรียมข้อมูล โดยเมื่อตรวจสอบข้อมูลแล้วพบว่าข้อมูลทั้งหมดเป็นข้อมูลที่ประกอบด้วย 649 รายการ 28 ตัวแปร ซึ่งแต่ละ ตัวแปร จะแสดงให้เห็นถึงคุณลักษณะเฉพาะของข้อมูล คือ integer แต่จะยกเว้นใน ตัวแปร G3 ซึ่งเป็น Polynominal และกำหนดให้เป็น Label เพราะเป็นข้อมูลที่ต้องการทำนาย ซึ่งแบ่งเป็น class pass 504 รายการ และ class fail 145 รายการ ดังภาพประกอบที่ 37

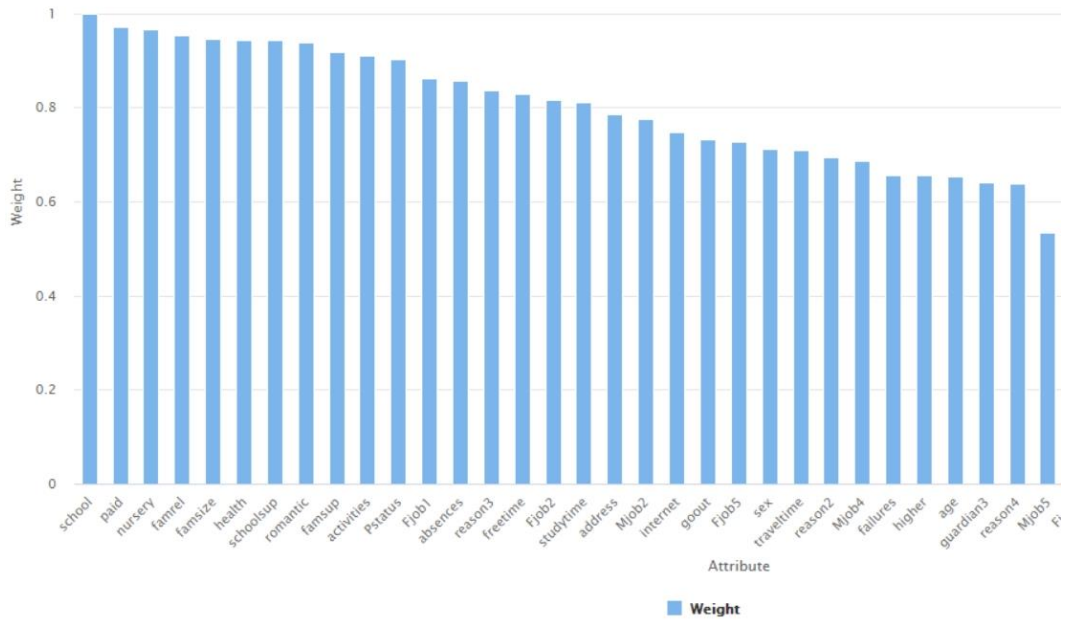
Name	Type	Missing	Statistics		
Label **G3**	Polynominal	0	Least fail (145)	Most pass (504)	Values pass (504), fail (145)
school	Integer	0	Min 1	Max 2	Average 1.348
sex	Integer	0	Min 1	Max 2	Average 1.410
age	Integer	0	Min 15	Max 22	Average 16.744

ภาพประกอบ 37 การตรวจสอบค่าต่าง ๆ ในข้อมูลทั้งหมด

เมื่อ Import ข้อมูลเข้าใน Rapid Miner โดยใช้ Operator Correlation Matrix ประมวลค่าน้ำหนัก Correlation ให้เป็นค่าตั้งแต่ 0 – 1 และ 0 – (-1) ดังภาพประกอบที่ 38

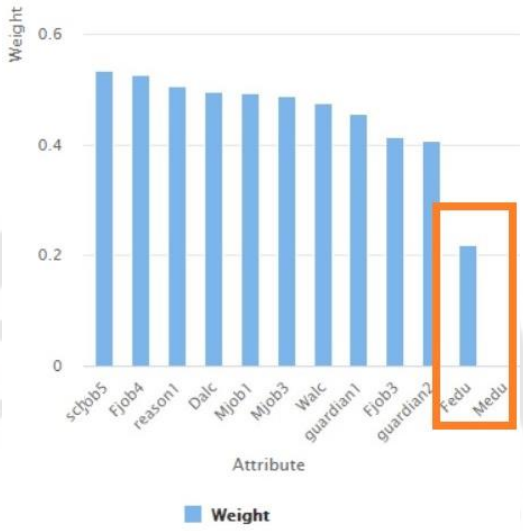
Attribut...	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob1	Mjob2	Mjob3	Mjob4	Mjob5	Fj
school	1	-0.024	0.119	0.024	0.058	-0.021	-0.029	-0.016	0.048	-0.071	0.008	-0.003	-0.011	0.
sex	-0.024	1	-0.044	-0.026	0.098	0.065	0.119	0.084	-0.134	0.040	-0.024	0.056	0.105	-0
age	0.119	-0.044	1	0.026	-0.002	-0.006	-0.108	-0.121	0.089	-0.100	0.039	-0.035	-0.047	0.
address	0.024	-0.026	0.026	1	-0.046	0.095	-0.190	-0.141	0.165	-0.084	0.039	-0.101	-0.073	-0
famsize	0.058	0.098	-0.002	-0.046	1	-0.240	-0.014	-0.040	0.009	0.010	-0.057	0.023	0.040	0.

ภาพประกอบ 38 ตารางแสดงค่า Correlation ของตัวแปร



ภาพประกอบ 39 การแสดงค่าน้ำหนักความสัมพันธ์ของแต่ละตัวแปรในชุดข้อมูล

จากภาพประกอบที่ 39 - 40 จะเห็นค่าความสัมพันธ์ของตัวแปรมีค่าระหว่าง 0 – 1 เมื่อนำข้อมูลมาหาความสัมพันธ์ของตัวแปร สำหรับงานวิจัยในครั้งนี้จะคัดเลือกเฉพาะตัวแปรที่มีค่าความสัมพันธ์ตั้งแต่ 0.4 ขึ้นไป ทำให้มีตัวแปร คงเหลือ 42 ตัวแปร ซึ่งตัวแปรที่ถูกตัดออก 2 ตัวแปร ได้แก่ Medu มีค่าความสัมพันธ์ เท่ากับ 0 และ Fedu มีค่าความสัมพันธ์ เท่ากับ 0.21 ดังภาพประกอบที่ 40



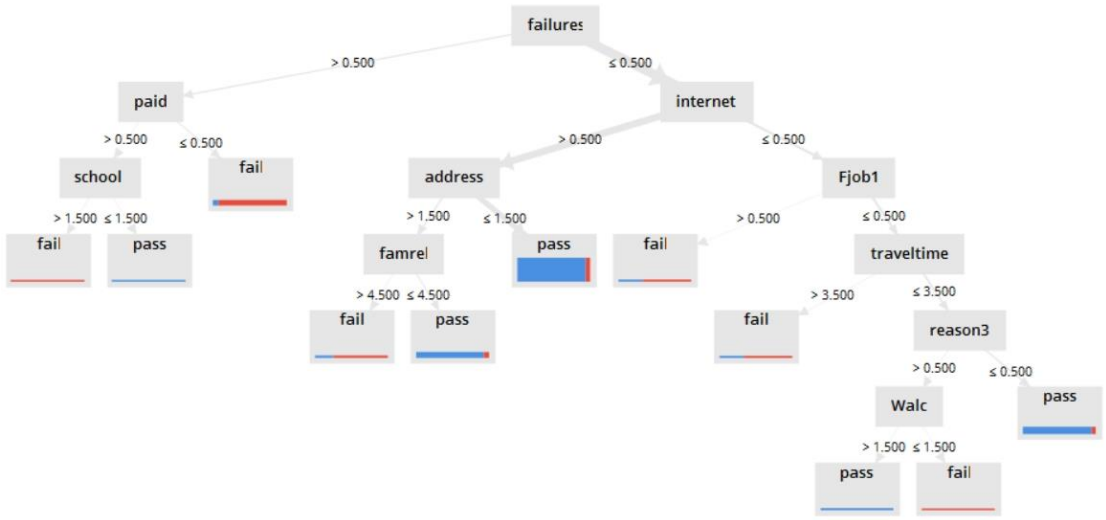
ภาพประกอบ 40 การแสดงค่าน้ำหนักความสัมพันธ์ของแต่ละตัวแปร Fedu และ Medu

4.2 ผลการศึกษาการสร้างตัวแบบทำนายการพยากรณ์ โดยใช้เทคนิคต่าง ๆ

เมื่อข้อมูลพร้อมสำหรับการทำนายแล้ว จึงนำมาสร้างตัวแบบทำนายและมีผลการสร้างตัวแบบทำนาย ดังนี้

4.2.1 เทคนิคต้นไม้ตัดสินใจ (Decision Tree)

ผู้วิจัยทำการสร้างตัวแบบทำนายด้วยเทคนิคต้นไม้ตัดสินใจ เพื่อทำการพยากรณ์ผลสัมฤทธิ์ทางการเรียนในเทอมที่ 3 จากนั้นผู้วิจัยได้นำข้อมูลเข้าสู่ตัวแบบทำนาย โดยซอฟต์แวร์ Rapid Miner นั้น Operator Decision Tree จะทำการวิเคราะห์คุณสมบัติที่ดีที่สุดของข้อมูลมาเป็นโหนดราก (Root node) และวนสร้างโหนดลูกและเส้นเชื่อมไปเรื่อย ๆ จนกว่าข้อมูลที่ได้จะถูกจัดไว้เป็นกลุ่มเดียวกันและหยุดสร้างต้นไม้



ภาพประกอบ 41 ตัวแบบทำนาย Decision Tree

จากภาพประกอบที่ 41 พบว่าตัวแบบทำนายมี 10 ขั้นตามที่กำหนด และมี failures เป็นโหนดราก ซึ่งเป็นตัวแปรที่สำคัญที่สุดในการทำนาย ต่อด้วยโหนดต่อไปจำนวน 10 โหนด โดยไล่ลำดับค่าที่มีค่าสูงเรียงลำดับไปเรื่อย ๆ จนครบ ซึ่งเชื่อมกันด้วยกิ่งที่แสดงค่าที่เป็นไปได้ของโหนด และโหนดใบแสดงคลาสที่กำหนดไว้ คือ pass และ fail ดังนี้

- เมื่อ failures และ paid มีค่ามากกว่า 0.5 school มีค่ามากกว่า 1.5 จะทำนายว่า fail แต่ถ้า school มีค่าน้อยกว่าหรือเท่ากับ 1.5 จะทำนายว่า pass
- เมื่อ failures มีค่าน้อยกว่า 0.5 และ paid มีค่ามากกว่าหรือเท่ากับ 0.5 จะทำนายว่า fail
- เมื่อ failures มีค่ามากกว่าหรือเท่ากับ 0.5 Internet มีค่ามากกว่า 0.5 address มีค่ามากกว่า 1.5 famrel มีค่ามากกว่า 4.5 จะทำนายว่า fail แต่ถ้า address มีค่าน้อยกว่าหรือเท่ากับ 1.5 จะทำนายว่า pass
- เมื่อ failures มีค่ามากกว่าหรือเท่ากับ 0.5 Internet มีค่าน้อยกว่าหรือเท่ากับ 0.5 Fjob1 มีค่ามากกว่า 0.5 จะทำนายว่า fail แต่ถ้า Fjob1 มีค่าน้อยกว่าหรือเท่ากับ 0.5 traveltime มีค่ามากกว่า 3.5 จะทำนายว่า fail แต่ถ้า traveltime มีค่าน้อยกว่าหรือเท่ากับ 3.5 และ reson3 มีค่าน้อยกว่าหรือเท่ากับ 0.5 จะทำนายว่า pass แต่ถ้า reson3 มีค่ามากกว่า 0.5

และ walc มีค่ามากกว่า 1.5 จะทำนายว่า pass แต่ถ้า walc มีค่าน้อยกว่าหรือเท่ากับ 1.5 จะทำนายว่า fail

เมื่อพิจารณาค่าในการทำนาย ซึ่งโมเดลได้ทำนายว่า จะมีเด็กที่สอบผ่าน 104 คน และสอบตก 26 คน แต่เมื่อเปรียบเทียบกับข้อมูลจริงพบว่า มีเด็กที่สอบผ่าน 101 คน และสอบตก 29 คน จึงมีค่าความถูกต้องในการพยากรณ์โมเดล(accuracy) เท่ากับ 91.54% ดังภาพประกอบที่ 42

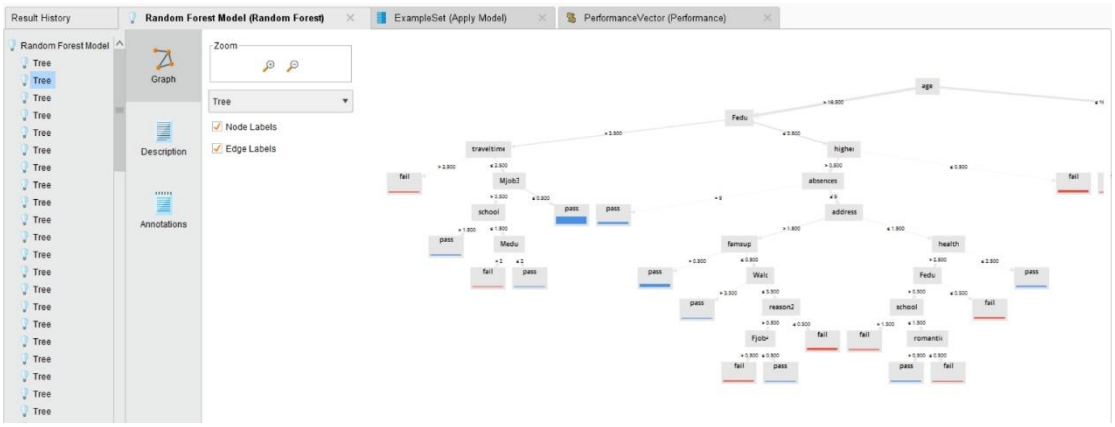
accuracy: 91.54%

	true pass	true fail	class precision
pred. pass	97	7	93.27%
pred. fail	4	22	84.62%
class recall	96.04%	75.86%	

ภาพประกอบ 42 ค่าความถูกต้องของการพยากรณ์ตัวแบบทำนาย Decision Tree

4.2.2 เทคนิคป่าแห่งการทำนาย (Random Forest)

ผู้วิจัยทำการสร้างตัวแบบทำนายด้วยเทคนิคป่าแห่งการทำนาย เพื่อทำการพยากรณ์ผลสัมฤทธิ์ทางการเรียนในเทอมที่ 3 โดยใช้ตัวแบบทำนายจากต้นไม้ตัดสินใจหลาย ๆ ต้น ตั้งแต่ 100 ตัวแบบทำนายขึ้นไป ซึ่งถูกสร้างจากการนำข้อมูลฝึกสอนไปสุ่มเลือกตัวอย่างข้อมูลและคุณลักษณะข้อมูลแล้วนำมาสร้างเป็นต้นไม้ตัดสินใจ จากนั้นผู้วิจัยได้นำข้อมูลเข้าสู่ตัวแบบทำนาย โดยซอฟต์แวร์ Rapid Miner นั้น Operator Random Forest จะทำการวิเคราะห์และใช้ตัวแบบทำนายจากต้นไม้ตัดสินใจ ดังภาพประกอบที่ 43 ทางด้านซ้ายของรูป คือ โมเดลต้นไม้ตัดสินใจ 100 ต้น



ภาพประกอบ 43 ตัวแบบทำนาย Random Forest

เมื่อพิจารณาค่าในการทำนาย ซึ่งโมเดลได้ทำนายว่า จะมีเด็กที่สอบผ่าน 101 คน และสอบตก 29 คน และเมื่อเปรียบเทียบกับข้อมูลจริงพบว่าการทำนายจำนวนเด็กที่สอบผ่าน และสอบตกนั้น ถูกต้องทั้งหมด จึงมีค่าความถูกต้องในการพยากรณ์โมเดล(accuracy) เท่ากับ 91.54% ดังภาพประกอบที่ 44

accuracy: 100.00%

	true pass	true fail	class precision
pred. pass	101	0	100.00%
pred. fail	0	29	100.00%
class recall	100.00%	100.00%	

ภาพประกอบ 44 ค่าความถูกต้องของการพยากรณ์ตัวแบบทำนาย Random Forest

4.1.3 เทคนิคการเรียนรู้เบย์ (Naïve-Bayes)

ผู้วิจัยทำการสร้างตัวแบบทำนายด้วยเทคนิคการเรียนรู้เบย์ เพื่อทำการพยากรณ์ผลสัมฤทธิ์ทางการเรียนในเทอมที่ 3 โดยการหาความน่าจะเป็นของสิ่งที่ยังไม่เคยเกิดขึ้นและคาดเดาจากสิ่งที่เคยเกิดขึ้นมาก่อน โดยซอฟต์แวร์ Rapid Miner นั้น Operator Naïve-Bayes จะทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปร ซึ่งไม่มีการปรับค่า parameter ใด

SimpleDistribution

```
Distribution model for label attribute G3


Class pass (0.777)
43 distributions

Class fail (0.223)
43 distributions
```

ภาพประกอบ 45 ตัวแบบทำนาย Naïve-Bayes

จากภาพประกอบที่ 45 พบว่า ค่าความน่าจะเป็นที่โมเดลทำนายทั้งหมด คือ ค่าความน่าจะเป็นไปของ Class pass เท่ากับ 0.777 และ ค่าความน่าจะเป็นไปของ Class fail เท่ากับ 0.223

เมื่อพิจารณาค่าในการทำนาย ซึ่งโมเดลได้ทำนายว่า จะมีเด็กที่สอบผ่าน 64 คน และสอบตก 66 คน แต่เมื่อเปรียบเทียบกับข้อมูลจริงพบว่า มีเด็กที่สอบผ่าน 101 คน และสอบตก 29 คน จึงมีค่าความถูกต้องในการพยากรณ์โมเดล(accuracy) เท่ากับ 91.54% ดังภาพประกอบที่ 46

accuracy: 68.46%

	true pass	true fail	class precision
pred. pass	62	2	96.88%
pred. fail	39	27	40.91%
class recall	61.39%	93.10%	

ภาพประกอบ 46 ค่าความถูกต้องของการพยากรณ์ตัวแบบทำนาย Naïve-Bayes

4.1.3 เทคนิคเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbors)

ผู้วิจัยทำการสร้างตัวแบบทำนายด้วยเทคนิคเพื่อนบ้านใกล้ที่สุด เพื่อทำการพยากรณ์ผลสัมฤทธิ์ทางการเรียนในเทอมที่ 3 โดยการค้นหาเพื่อนบ้านใกล้ที่สุด เป็นการแบ่งกลุ่มข้อมูล และวัดระยะห่างระหว่างข้อมูลที่ต้องการทำนายกับข้อมูลที่อยู่ใกล้เคียงเป็นจำนวน K ตัว ซึ่งซอฟต์แวร์ Rapid Miner นั้น Operator K-Nearest Neighbors จะทำการวิเคราะห์ข้อมูลหลังจากนั้นดูค่า “k” ว่ากำหนดเท่าใด แล้วทำการสร้างแบบทำนาย

KNNClassification

```
Weighted 5-Nearest Neighbour model for classification.
The model contains 130 examples with 43 dimensions of the following classes:
  pass
  fail
```

ภาพประกอบ 47 ตัวแบบทำนาย k-NN

จากภาพประกอบที่ 47 กำหนดให้ค่า k = 5 โดยมีข้อมูลทั้งหมด 130 รายการ 43 คอลัมน์ ซึ่งทำนายหาว่าแต่ละรายการจะเข้าใกล้ และเป็นคลาสใด ระหว่าง pass และ fail

เมื่อพิจารณาค่าในการทำนาย ซึ่งโมเดลได้ทำนายว่า จะมีเด็กที่สอบผ่าน 120 คน และสอบตก 10 คน แต่เมื่อเปรียบเทียบกับข้อมูลจริงพบว่า มีเด็กที่สอบผ่าน 101 คน และสอบตก 29 คน จึงมีค่าความถูกต้องในการพยากรณ์โมเดล(accuracy) เท่ากับ 82.31% ดังภาพประกอบที่ 48

accuracy: 82.31%

	true pass	true fail	class precision
pred. pass	99	21	82.50%
pred. fail	2	8	80.00%
class recall	98.02%	27.59%	

ภาพประกอบ 48 ค่าความถูกต้องของการพยากรณ์ตัวแบบทำนาย k-NN

จากผลการศึกษาการสร้างตัวแบบทำนายการพยากรณ์โดยใช้เทคนิคต่าง ๆ จะเห็นได้ว่า เทคนิคป่าแห่งการทำนาย (Random Forest) ให้ค่าความถูกต้อง (accuracy) มากที่สุด เท่ากับ 100% และเทคนิคการเรียนรู้เบย์ (Naïve-Bayes) ให้ค่าความถูกต้อง (accuracy) น้อยที่สุดเท่ากับ 68.46% ดังตารางที่ 14

ตาราง 14 ผลการเปรียบเทียบค่าความถูกต้องจากการสร้างตัวแบบทำนาย

เทคนิคเหมืองข้อมูล	ค่าความถูกต้อง %
เทคนิคต้นไม้ตัดสินใจ (Decision Tree)	91.54
เทคนิคป่าแห่งการทำนาย (Random Forest)	100
เทคนิคการเรียนรู้เบย์ (Naïve-Bayes)	68.46
เทคนิคเพื่อนบ้านใกล้ที่สุด (K-NN)	82.31

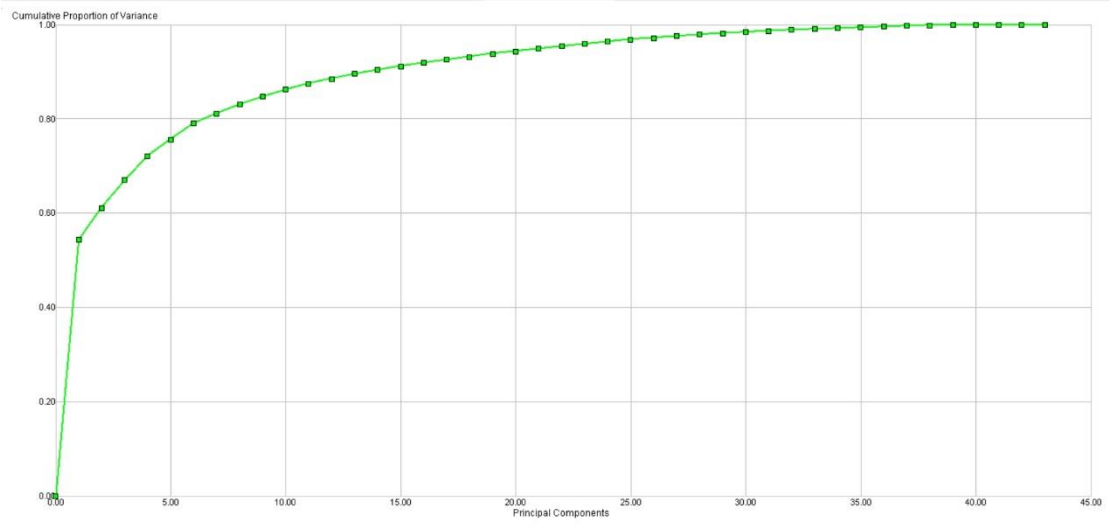
4.2 ผลการศึกษาการสร้างตัวแบบทำนายการพยากรณ์ด้วยเทคนิคการวิเคราะห์องค์ประกอบหลักของข้อมูล

เนื่องจากเมื่อข้อมูลได้ผ่านขั้นตอนการเตรียมข้อมูลแล้ว ทำให้มีตัวแปรทั้งหมด 42 ตัวแปร ซึ่งทางผู้วิจัยยังไม่ทราบแน่ชัดว่าข้อมูลใดบ้างที่จะมีคุณลักษณะที่จะสามารถใช้งานได้ อาจส่งผลต่อความถูกต้องของการประมวลผล ดังนั้นผู้วิจัยจึงทำการวิเคราะห์องค์ประกอบหลักของข้อมูล โดยใช้วิธี Principal Component Analysis (PCA) เพื่อย่อย วิเคราะห์ และหาความสัมพันธ์ของข้อมูลเหล่านั้น ให้มีความกระชับมากขึ้น โดยไม่ทำให้สูญเสียข้อมูลสำคัญไป ส่งผลให้มีตัวแปรคงเหลือ 22 ตัวแปร ดังภาพประกอบที่ 49 ซึ่งใช้สำหรับการสร้างตัวแบบทำนาย

pc_9	pc_10	pc_11	pc_12	pc_13	pc_14	pc_15	pc_16	pc_17	pc_18	pc_19	pc_20	pc_21	pc_22
0.400	-0.314	-1.388	0.612	-0.611	0.694	-0.195	0.409	0.148	0.245	-0.258	0.311	-0.631	-0.184
1.565	2.057	0.756	-0.582	1.112	-0.686	0.032	-0.366	-0.490	0.994	-0.165	0.250	-0.168	0.684
0.144	-0.203	0.057	-0.102	0.073	0.548	0.170	-0.486	-0.243	0.061	0.905	-0.385	-0.002	-0.384
1.770	-0.054	-0.281	-0.357	0.257	0.001	-0.891	0.190	0.295	-0.288	-0.048	-0.201	-0.523	0.796

ภาพประกอบ 49 ตัวอย่างตัวแปรที่ผ่านกระบวนการ PCA

จากภาพประกอบที่ 50 จะแสดงค่า Variance ของตัวแปร และจะสังเกตได้ว่าตัวแปรที่มีค่า Variance ช่วง 0.95 มีทั้งหมด 22 ตัว



ภาพประกอบ 50 การแสดงค่า Variance ตัวแปรเมื่อผ่านกระบวนการ PCA

โดยผู้วิจัยได้นำข้อมูลทั้งหมดที่ผ่านกระบวนการวิเคราะห์องค์ประกอบหลักของข้อมูล (PCA) แล้วนั้นเข้าสู่ตัวแบบทำนาย โดยใช้เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เทคนิคป่าแห่งการทำนาย (Random Forest) เทคนิคการเรียนรู้เบย์ (Naïve-Bayes) เทคนิคเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbors) เพื่อพยากรณ์ผลการเรียนในเทอมที่ 3 พบว่าเทคนิคป่าแห่งการทำนาย (Random Forest) ให้ค่าความถูกต้อง (accuracy) มากที่สุด เท่ากับ 99.23% แต่ลดลงเมื่อเทียบผลลัพท์ที่ไม่ใช้ PCA และเทคนิคเพื่อนบ้านใกล้ที่สุด (K-NN) ให้ค่าความถูกต้อง (accuracy) น้อยที่สุดเท่ากับ 82.31% ซึ่งมีผลลัพท์เท่าเดิมเมื่อเทียบกับการไม่ใช้ PCA ส่วนเทคนิคป่าแห่งการทำนาย (Random Forest) และเทคนิคการเรียนรู้เบย์ (Naïve-Bayes)) ให้ค่าความถูกต้อง (accuracy) มากขึ้นเมื่อเทียบกับผลลัพท์ที่ไม่ใช้ PCA ดังตารางที่ 15

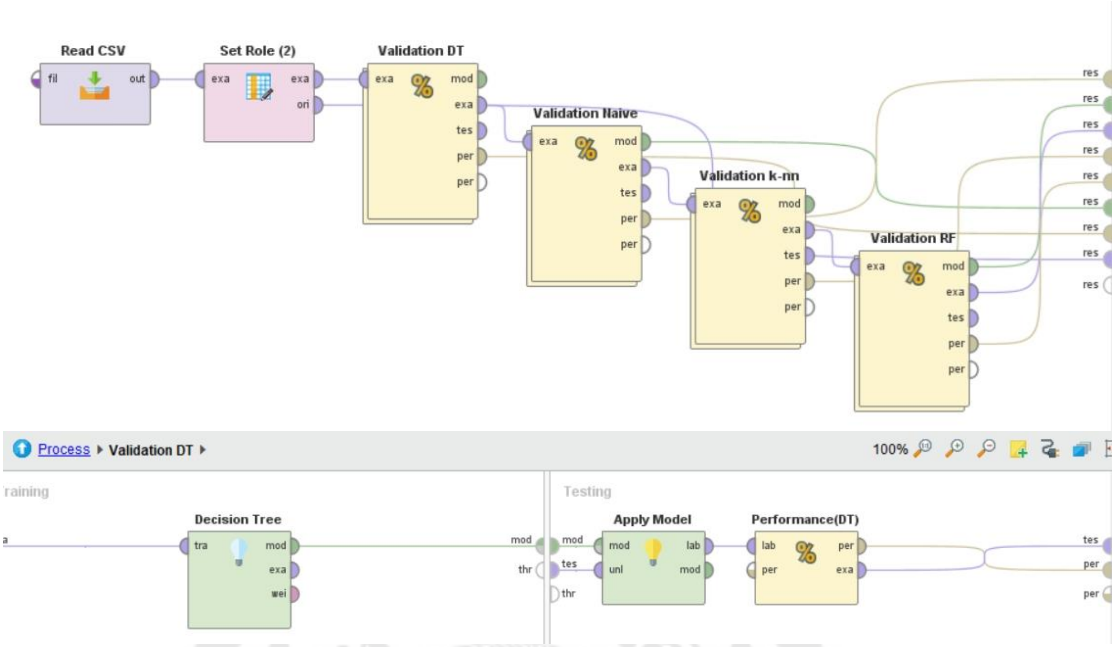
ตาราง 15 ผลการเปรียบเทียบค่าความถูกต้องจากการสร้างตัวแบบทำนาย เมื่อข้อมูลผ่าน PCA

เทคนิคเหมืองข้อมูล	ค่าความถูกต้อง %
เทคนิคต้นไม้ตัดสินใจ (Decision Tree)	94.62
เทคนิคป่าแห่งการทำนาย (Random Forest)	99.23
เทคนิคการเรียนรู้เบย์ (Naïve-Bayes)	85.38
เทคนิคเพื่อนบ้านใกล้ที่สุด (K-NN)	82.31

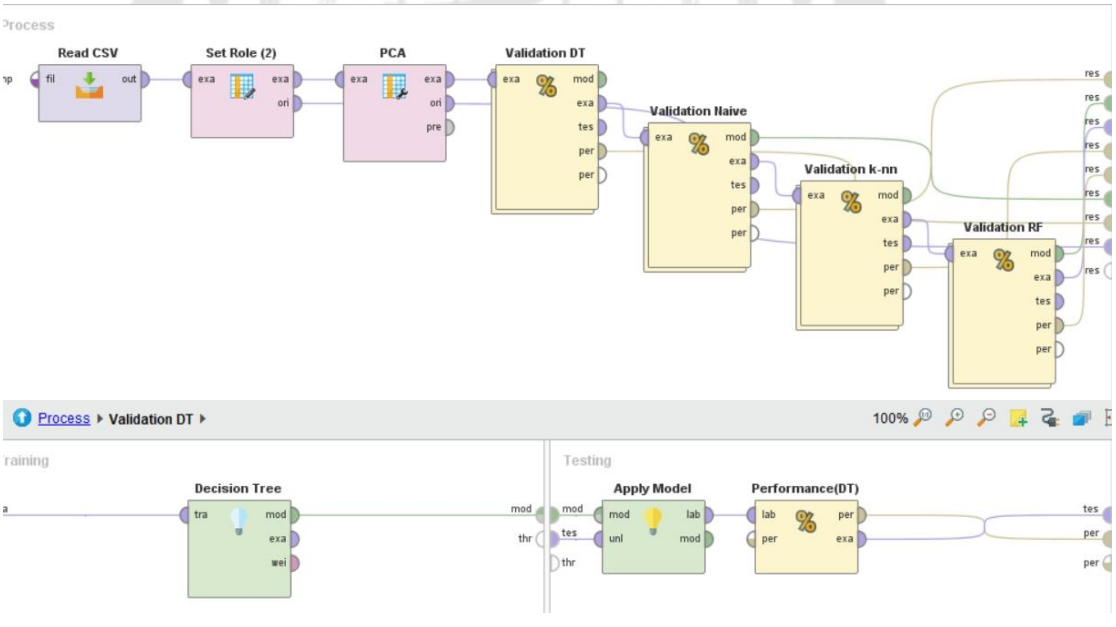
4.3 ผลการวิเคราะห์ความแม่นตรง และเปรียบเทียบประสิทธิภาพของตัวแบบทำนายด้วยเทคนิคต่าง ๆ โดยใช้วิธี K-fold cross validation

หลังจากการสร้างตัวแบบทำนายการพยากรณ์ด้วยเทคนิคทั้ง 4 เทคนิคข้างต้นแล้วนั้น ผู้วิจัยได้ทำการวิเคราะห์ความแม่นตรงของตัวแบบทำนาย โดยใช้วิธี K-fold cross validation คือ การแบ่งข้อมูลเป็น k ส่วนเท่า ๆ กันเพื่อสร้างและทดสอบตัวแบบทำนาย

โดยในงานวิจัยนี้ได้แบ่งข้อมูลเป็น 10 ส่วนเท่า ๆ กัน และทำการทดสอบ ซึ่งข้อมูลที่ใช้ในการสร้างตัวแบบทำนายการพยากรณ์จะทดลองใช้ข้อมูลทั้งที่ไม่ผ่านกระบวนการวิเคราะห์องค์ประกอบหลักของข้อมูล (PCA) ดังภาพประกอบที่ 51 และผ่านกระบวนการวิเคราะห์องค์ประกอบหลักของข้อมูล (PCA) ดังภาพประกอบที่ 52 นำเข้าสู่โอเปอร์เรเตอร์ Cross validation โดยกำหนดให้แบ่งข้อมูลแบบ Stratified sampling เพื่อให้ได้จำนวนกลุ่มตัวอย่างตามสัดส่วนของขนาดกลุ่มตัวอย่างและกลุ่มประชากร ในการหาค่าความถูกต้อง (accuracy) ค่าความแม่นยำ (precision) ค่าความระลึก (recall) และค่าความถ่วงดุล (f1_measure)



ภาพประกอบ 51 การวิเคราะห์ความแม่นตรงโดยใช้ข้อมูลที่ไม่ผ่าน PCA



ภาพประกอบ 52 การวิเคราะห์ความแม่นตรงโดยใช้ข้อมูลที่ผ่าน PCA

จากการศึกษาการสร้างตัวแบบทำนายการพยากรณ์ด้วยเทคนิคทั้ง ด้วยเทคนิคทั้ง 4 เทคนิคข้างต้น และทำการวิเคราะห์ความแม่นตรงของตัวแบบทำนาย ดังตารางที่ 16 พบว่าข้อมูลที่ไม่ผ่านกระบวนการวิเคราะห์องค์ประกอบหลักของข้อมูล (PCA) นั้น ซึ่งผู้วิจัยใช้ตัวย่อแทนข้อมูล

นี้ว่า Ori เทคนิคป่าไม้ตัดสินใจ (Random Forest) ให้ค่าความถูกต้อง (accuracy) มากที่สุด เท่ากับ 80.58% ให้ค่าความแม่นยำ (precision) มากที่สุด เท่ากับ 66.49% ให้ค่าความระลึก (recall) เป็นอันดับ 3 เท่ากับ 28.38% และให้ค่าความถ่วงดุล (f1_measure) เป็นอันดับ 3 เท่ากับ 39.78% ซึ่งเทคนิคต้นไม้ตัดสินใจ (Decision Tree) ให้ค่าความถูกต้อง (accuracy) น้อยที่สุด เท่ากับ 76.28% ให้ค่าความแม่นยำ (precision) น้อยที่สุด เท่ากับ 33.02% ให้ค่าความระลึก (recall) เป็นอันดับ 2 เท่ากับ 37.43% และให้ค่าความถ่วงดุล (f1_measure) เป็นอันดับ 2 เท่ากับ 41.59%

ข้อมูลที่ผ่านกระบวนการวิเคราะห์องค์ประกอบหลักของข้อมูล (PCA) นั้น เทคนิคป่าไม้ตัดสินใจ (Random Forest) ให้ค่าความถูกต้อง (accuracy) มากที่สุด เท่ากับ 80.74% ให้ค่าความแม่นยำ (precision) มากที่สุด เท่ากับ 76.50% ให้ค่าความระลึก (recall) เป็นอันดับ 3 เท่ากับ 19.57% และให้ค่าความถ่วงดุล (f1_measure) เป็นอันดับ 2เท่ากับเท่ากับ 31.16% ซึ่งเทคนิคต้นไม้ตัดสินใจ (Decision Tree) ให้ค่าความถูกต้อง (accuracy) น้อยที่สุด เท่ากับ 76.28% ให้ค่าความแม่นยำ (precision) น้อยที่สุดเท่ากับ 33.02% ให้ค่าความระลึก (recall) เป็นอันดับ 2 เท่ากับ 25.10% และให้ค่าความถ่วงดุล (f1_measure) เป็นอันดับ 2 เท่ากับ 28.52%

ตาราง 16 ผลการเปรียบเทียบประสิทธิภาพจากการสร้างตัวแบบทำนาย โดยใช้วิธี K-fold cross validation

เทคนิคเหมืองข้อมูล	ค่าความถูกต้อง %		ค่าความแม่นยำ %		ค่าความระลึก %		ค่าความถ่วงดุล %	
	Ori	PCA	Ori	PCA	Ori	PCA	Ori	PCA
เทคนิคต้นไม้ตัดสินใจ (Decision Tree)	76.28	71.35	46.81	33.02	37.43	25.10	41.59	28.52
เทคนิคป่าไม้ตัดสินใจ (Random Forest)	**80.58**	**80.74**	66.49	76.50	28.38	19.57	39.78	31.16
เทคนิคการเรียนรู้เบย์ (Naïve-Bayes)	76.58	78.28	**50.19**	**55.09**	**62.14**	**30.48**	**55.52**	**39.24**
เทคนิคเพื่อนบ้านใกล้ที่สุด (K-NN)	78.43	78.43	61.90	60.00	9.00	10.38	15.71	17.6

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

จากการศึกษาการสร้างตัวแบบทำนายผลสัมฤทธิ์ทางการเรียนในเทอมที่ 3 ของนักเรียนทั้งหมด 649 คน ตัวแปรทั้งหมด 33 ตัวแปร ซึ่งต้องการทำนายว่า นักเรียนจะสอบผ่าน (pass) หรือสอบตก(fail) โดยใช้โปรแกรม RapidMiner studio 9.5 เป็นซอฟต์แวร์ในการวิเคราะห์ด้วยเทคนิคเหมืองข้อมูลต่าง ๆ ซึ่งก่อนการสร้างตัวแบบทำนายผู้วิจัยได้ทำตามกระบวนการทำเหมืองข้อมูลแบบ CRISP-DM โดยจัดเตรียมข้อมูลให้พร้อม ซึ่งต้องทำการแปลงข้อมูล ให้อยู่ในรูปแบบของตัวเลขทั้งหมด และหาความสัมพันธ์ของข้อมูลก่อนการทำนาย (Correlation) จึงทำให้สามารถคัดเลือกตัวแปรได้ทั้งหมด 42 ตัวแปร และได้ศึกษาและเลือกเทคนิคที่ใช้สร้างตัวแบบทำนาย 4 เทคนิค คือ เทคนิคต้นไม้ตัดสินใจ (Decision Tree), เทคนิคป่าไม้ตัดสินใจ (Random Forest), เทคนิคการเรียนรู้เบย์ (Naïve-Bayes), เทคนิคเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbors) รวมทั้งการวิเคราะห์องค์ประกอบหลักของข้อมูลที่ส่งผลต่อประสิทธิภาพของตัวแบบทำนาย และการวิเคราะห์ความแม่นตรงของโมเดล K-fold Cross Validation จากนั้นจึงเริ่มศึกษาการสร้างแบบจำลองด้วยเทคนิคต่าง ๆ โดยก่อนการสร้างตัวแบบทำนาย ผู้วิจัยได้แบ่งข้อมูลเป็น 2 กลุ่ม คือ 80 : 20 และการศึกษาวิจัยมีการปรับค่าตัวแปรที่เกี่ยวข้องเพื่อให้แบบทำนายให้ผลลัพธ์ที่มีประสิทธิภาพที่ดีที่สุด ซึ่งมีรายละเอียดดังนี้

1. การศึกษาการสร้างตัวแบบทำนายการพยากรณ์ โดยเทคนิคต่าง ๆ โดยมีการใช้ เทคนิคต้นไม้ตัดสินใจ (Decision Tree) โดยผู้ศึกษาวิจัยปรับเกณฑ์การวัดผลเป็น gini index เพื่อเลือกตัวแปรที่ดีที่สุดที่สามารถแยกความแตกต่างระหว่างเรคคอร์ดต่าง ๆ ในชุดข้อมูลได้ และ กำหนดค่า maximal depth เท่ากับ 10 พบว่าค่าความถูกต้องของการพยากรณ์ (accuracy) เท่ากับ 91.54% ใช้เทคนิคป่าไม้ตัดสินใจ (Random Forest) โดยผู้ศึกษาวิจัยปรับค่าจำนวนของ tree ที่นำมาใช้ในการสร้างตัวแบบการทำนายที่ 100 ต้น และกำหนดค่า maximal depth ของแต่ละ tree ที่ 15 พบว่าค่าความถูกต้องของการพยากรณ์ (accuracy) เท่ากับ 100 % ซึ่งเป็นค่าที่ดีที่สุด การใช้เทคนิคการเรียนรู้เบย์ (Naïve-Bayes) โดยผู้วิจัยได้นำข้อมูลเข้าสู่ตัวแบบทำนายเข้าสู่ซอฟต์แวร์ Rapid Miner ซึ่งไม่มีการปรับค่า parameter ใด พบว่าค่าความถูกต้องของการพยากรณ์ (accuracy) เท่ากับ 68.46% ซึ่งเป็นค่าที่ดีที่น้อยที่สุด การใช้เทคนิคเทคนิคเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbors) โดยผู้วิจัยได้นำข้อมูลเข้าสู่ตัวแบบทำนาย กำหนดค่า “k” เท่ากับ 5 ที่ทำ

ให้ค่าความถูกต้องดีที่สุด และพบว่าค่าความถูกต้องของการพยากรณ์ (accuracy) เท่ากับ 82.31%

จะเห็นได้ว่าผลการศึกษาการสร้างตัวแบบทำนายการพยากรณ์ โดยใช้เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เทคนิคป่าแห่งการทำนาย (Random Forest) และเทคนิคเพื่อนบ้านใกล้ที่สุด (K-NN) ให้ค่าความถูกต้อง (accuracy) อยู่ในระดับมาก ส่วนเทคนิคการเรียนรู้เบย์ (Naïve-Bayes) ให้ค่าความถูกต้อง (accuracy) อยู่ในระดับน้อยที่สุด ซึ่งมีความเป็นไปได้ว่าการสร้างตัวแบบทำนายโดยซอฟต์แวร์ Rapid Miner ใน 3 เทคนิคที่ให้ค่าสูง สามารถปรับแต่งค่าพารามิเตอร์ในแต่ละตัวแบบทำนาย เพื่อให้ได้ตัวแบบทำนายที่เหมาะสมที่สุด ในขณะที่เทคนิคการเรียนรู้เบย์ (Naïve-Bayes) ไม่สามารถปรับแต่งค่าใดได้

2. การวิเคราะห์องค์ประกอบหลักของข้อมูลที่ส่งผลต่อประสิทธิภาพของตัวแบบทำนาย เมื่อข้อมูลได้ผ่านขั้นตอนการเตรียมข้อมูลแล้ว ทำให้จำนวนตัวแปรที่ใช้มีจำนวนเพิ่มขึ้นเท่ากับ 42 ตัวแปร โดยผู้ศึกษาวิจัยจึงได้นำเทคนิคการวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis : PCA) เพื่อทำการวิเคราะห์และคัดเลือกตัวแปรเพื่อนำไปสร้างตัวแบบทำนายการพยากรณ์อีกครั้ง ซึ่งทำให้มีตัวแปรลดลงเท่ากับ 22 ตัวแปร ซึ่งพบว่ากระบวนการวิเคราะห์องค์ประกอบหลักของข้อมูล (PCA) ที่เลือกใช้นั้น ส่งผลทำให้ค่าความถูกต้อง (accuracy) ของเทคนิคต้นไม้ตัดสินใจ (Decision Tree) และเทคนิคการเรียนรู้เบย์ (Naïve-Bayes) เพิ่มขึ้น ในขณะที่เทคนิคเพื่อนบ้านใกล้ที่สุด (K-NN) เท่าเดิม และ เทคนิคป่าแห่งการทำนาย (Random Forest) ลดลง

จะเห็นได้ว่าเทคนิคการวิเคราะห์องค์ประกอบหลักของข้อมูลส่งผลต่อค่าความถูกต้อง (accuracy) ของตัวแบบทำนาย ซึ่งส่งผลให้ตัวแบบทำนายมีประสิทธิภาพเพิ่มขึ้นแค่เฉพาะบางเทคนิค และบางเทคนิคกลับมีประสิทธิภาพลดลง นั่นหมายความว่า เทคนิคการสร้างตัวแบบทำนายที่เลือกใช้นั้น บางเทคนิคอาจจะเหมาะสมกับการใช้ตัวแปรที่มีจำนวนมาก และมีความซับซ้อน เมื่อผู้วิจัยทำการลดจำนวน และขนาดความซับซ้อนลง จึงส่งผลต่อค่าความถูกต้อง (accuracy) ให้ลดลงด้วย

3. การวิเคราะห์ความแม่นตรงของโมเดล K-fold Cross Validation เพื่อตรวจสอบค่าความผิดพลาดในการพยากรณ์ของตัวแบบทำนาย โดยผู้ศึกษาวิจัยกำหนดให้ค่า k เท่ากับ 10 และทำการสร้างตัวแบบทำนาย ซึ่งจะใช้ข้อมูลทั้งที่ไม่ผ่าน (Ori) และผ่าน (PCA) การวิเคราะห์องค์ประกอบหลักของข้อมูล พบว่าค่าความถูกต้อง(accuracy) โดยรวมของการทำนายเมื่อเทียบกับข้อมูลที่เราใช้ในการทำนายนักเรียนที่มีผลการเรียนในเทอมที่ 3 เป็น pass และ fail พบว่า

เทคนิคที่มีความถูกต้องมากที่สุด คือ เทคนิคป่าไม้ตัดสินใจ (Random Forest) เท่ากับ 80.58(Ori) และ 80.74(PCA)

จากการทดลองจะเห็นได้ว่าการสร้างตัวแบบทำนายผลสัมฤทธิ์ทางการเรียนในเทอมที่ 3 ด้วยเทคนิคเหมืองข้อมูลทั้ง 4 เทคนิคให้ผลลัพธ์ที่แตกต่างกัน จึงทำให้ผู้วิจัยทราบว่าลักษณะของข้อมูลที่ใช้ในการสร้างแบบจำลองมีผลต่อการทำนาย และการปรับค่าพารามิเตอร์ ก็ส่งผลให้ประสิทธิภาพของการทำนายเพิ่มขึ้นหรือลดลงได้เช่นกัน ทั้งนี้เทคนิคการวิเคราะห์องค์ประกอบหลักของข้อมูล ถึงแม้จะทำให้จำนวนตัวแปรที่มีจำนวนมากสามารถลดจำนวนลงได้ แต่ก็ยังไม่สามารถส่งผลที่เป็นที่น่าพอใจเมื่อนำมาใช้ และไม่สามารถดูได้ว่าตัวแปรใดที่ส่งผลให้ตัวแบบทำนายมีประสิทธิภาพสูงสุด นอกจากการลดจำนวนของตัวแปรลงเท่านั้น แต่การใช้ซอฟต์แวร์ Rapid Miner นั้น ช่วยให้ผู้วิจัยสามารถสร้างแบบจำลองได้ง่าย โดยที่ไม่ต้องทำการ Coding อีกทั้งยังใช้เวลาในการประมวลได้อย่างรวดเร็ว

4. การเปรียบเทียบความเหมาะสมในการเลือกเทคนิคที่ใช้ในการสร้างตัวแบบพยากรณ์ เนื่องจากงานวิจัยนี้เป็นการทำนายผลการสอบของเด็กในเทอมที่ 3 และเมื่อพิจารณาจากสิ่งที่เราสนใจ คือ การทำนายเด็กที่สอบตกจริงว่าสอบตก ซึ่งเป็นค่า True Positive และ การทำนายเด็กที่สอบตกจริงว่าสอบผ่าน ซึ่งเป็นค่า False Positive นั้น จะเป็นค่าที่มีประโยชน์กับคุณครูมากที่สุด ซึ่งจะช่วยให้คุณครูให้ความสนใจกับเด็กคนดังกล่าวได้และช่วยให้เด็กนักเรียนกลุ่มนั้น ๆ สามารถมีผลสัมฤทธิ์ในเทอมที่ 3 เป็น ผ่าน ได้ในที่สุด

โดยเมื่อพิจารณาจากค่า True Positive ที่เราต้องการให้ค่านี้อยู่ในระดับที่สูง ดังนั้นเราจึงต้องการให้ค่า ค่าความระลึก (Recall) อยู่ในระดับที่สูงด้วย และเมื่อพิจารณาจากค่า False Positive ที่เราต้องการให้ค่านี้อยู่ในระดับที่ต่ำ ดังนั้นเราจึงต้องการให้ค่า ค่าความแม่นยำ (Precision) อยู่ในระดับที่ต่ำด้วย ซึ่งพบว่า เทคนิคการเรียนรู้เบย์ (Naïve-Bayes) ให้ค่าความระลึก (Recall) อยู่ในระดับที่สูงที่สุด เท่ากับ 62.14(Ori) และ 30.48 (PCA) และให้ค่าความแม่นยำ (Precision) อยู่ในระดับที่ต่ำ เท่ากับ 50.19(Ori) และ 55.09 (PCA) จึงทำให้ค่าความถ่วงดุล (f1_measure) ของเทคนิคการเรียนรู้เบย์ (Naïve-Bayes) เป็นเทคนิคที่มีค่าเฉลี่ยมากที่สุด ดังนั้นจึงเลือก เทคนิคการเรียนรู้เบย์ (Naïve-Bayes) เป็นเทคนิคที่ใช้ในการสร้างตัวแบบพยากรณ์ผลสัมฤทธิ์ในเทอมที่ 3 ของนักเรียน

5.2 ข้อเสนอแนะ

การเตรียมข้อมูลก่อนการทำการวิเคราะห์และสร้างแบบทำนายนั้น มีความสำคัญเป็นอย่างมาก เพราะหากทำการแปลงข้อมูลผิดพลาด อาจส่งผลต่อตัวแบบทำนาย รวมทั้งเมื่อทำการแปลงข้อมูลแล้วส่งผลทำให้ได้ข้อมูลที่มีจำนวนมากขึ้น สำหรับงานวิจัยนี้ในอนาคต อาจใช้วิธีการคัดเลือกคุณลักษณะของข้อมูลวิธีอื่น ๆ ที่จะสามารถลดมิติของข้อมูล หรือลดจำนวนของข้อมูลลงได้ เช่น การคัดเลือกคุณลักษณะแบบแรปเปอร์ ซึ่ง Ferda Ünal ได้ใช้ในการทำนายผลการเรียนของเด็กนักเรียน รัชพล กลัดชื่น และ จรัญ แสนราช (2561) ก็ยังใช้การคัดเลือกคุณลักษณะแบบฟิลเตอร์ และการคัดเลือกคุณลักษณะแบบฝังตัว ซึ่งช่วยลดมิติของข้อมูลเพื่อใช้ในการทำนายผลสัมฤทธิ์ทางการเรียนของนักศึกษาระดับอาชีวศึกษา และเห็นถึงความแตกต่างของค่าความถูกต้องในการสร้างตัวแบบทำนายอย่างชัดเจนมากขึ้น อีกทั้งการแบ่งข้อมูลในการสร้างตัวแบบทำนาย

ทั้งนี้การสุ่มหรือแบ่งข้อมูลก่อนการสร้างตัวแบบทำนายก็มีความสำคัญที่จะทำให้ข้อมูลของเราไม่มีความโอนเอียงไปข้างใดข้างหนึ่งจนเกินไป ดังนั้นขั้นตอนการแบ่งข้อมูลเพื่อใช้เป็นกลุ่มข้อมูลสอน (Training Data) และกลุ่มข้อมูลทดสอบ (Test Data) ควรต้องคำนึงถึงข้อมูลโดยต้องกำหนดให้ข้อมูลของเด็กที่สอบผ่านและสอบตก อยู่ในจำนวนที่เหมาะสม เพื่อไม่ให้โมเดลมีข้อมูลใดข้อมูลหนึ่งมากเกินไป และทำนายได้ถูกต้องมากยิ่งขึ้น และนอกจากนี้ยังมีเทคนิคเหมืองข้อมูลอื่น ๆ ที่สามารถสร้างตัวแบบทำนายเพื่อพยากรณ์ผลการเรียนของนักเรียนจาก Data set นี้ได้

บรรณานุกรม

Cortez, P. (2008). Student Performance Data Set. Retrieved from https://archive.ics.uci.edu/ml/datasets/Student+Performance

Cortez, P., & Silva, A. (2008). *Using Data Mining to Predict Secondary School Student Performance.* Paper presented at the FUture BUsiness TEChnology Conference (FUBUTEC 2008)

Cube, D. (2564). เอกสารประกอบการอบรมหลักสูตรอบรม Practical Data Mining with RapidMiner Studio 9. Retrieved from สืบค้นเมื่อวันที่ 30 มกราคม 2564, จาก https://datacubeth.ai/crisp-dm/

Deepika, K., & Sathyanarayana, N. (2018). Comparison Of Student Academic Performance On Different Educational Dataset Using Different Data Mining Techniques. *International Journal of Computational Engineering Research (IJCER),* 08(09), 28-38.

LTD, C. C. (2560). สถิติเบื้องต้นง่าย ๆ ที่จะทำให้คุณเข้าใจการวิเคราะห์มากขึ้น. Retrieved from สืบค้นเมื่อวันที่ 30 มกราคม 2564, จาก https://medium.com/@info_46914/

Petkovic, D., & Yang, S. H. B. J. From Explaning How Random Forest Classifier Predicts Learning of Software Engineering Teamwork to Guidance for Education.

Ünal, F. (2020). Data mining for student performance prediction in education. Retrieved from https://www.intechopen.com/online-first/data-mining-for-student-performance-prediction-in-education

ชณิดาภา บุญประสม, & จรัญ แสนราช. (2561). การวิเคราะห์การทำนายการลาออกกลางคันของนักศึกษา ระดับปริญญาตรีโดยใช้เทคนิควิธีการทำเหมืองข้อมูล. วารสารวิชาการครุศาสตร์อุตสาหกรรม พระจอมเกล้าพระนครเหนือ, 9(1), 98-105.

ชิตพงษ์ กิตตินราดร. (2562). Categorical Encoding. Retrieved from สืบค้นเมื่อวันที่ 30 มกราคม 2564, จาก https://guopai.github.io/ml-blog05.html

ณัฐพร เห็นเจริญเลิศ. (2558). รายงานผลการเข้าฝึกอบรมหลักสูตร การวิเคราะห์ข้อมูลด้วยเทคนิค *Data Mining* โดยซอฟต์แวร์ *RapidMiner Studio* 6 (ขั้นพื้นฐานและปานกลาง). Retrieved

from สาขาวิชาวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยเกษตรศาสตร์วิทยาเขตบางเขน : กรุงเทพฯ:

ธีระดา ภิญโญ. (2561). เทคนิคการแปลผลการวิเคราะห์องค์ประกอบสำหรับงานวิจัย. วารสารปัญญาภิวัฒน์(10), 290-304.

ปรเมษฐ์ ธันวานนท์, ชัยกรยิ่ง เสรี, วรพล พงษ์เพ็ชร, & และธนภัทร ฆังคะจิตรสกสรรค์. (2560). การประยุกต์ใช้โมเดลการเรียนรู้แบบรวมกลุ่มเพื่อพยากรณ์แนวโน้มของราคาหลักทรัพย์ในตลาดหลักทรัพย์แห่งประเทศไทย. *JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY,* 7, 12-21.

ปริวรรต เพียรภายลุน, ธาดา จันตะคุณ, & รักถิ่น เหลาหา. (2560). การเปรียบเทียบตัวแบบสำหรับใช้พยากรณ์คะแนน *O-NET* ด้วยเทคนิคเหมืองข้อมูล. Paper presented at the การประชุมวิชาการระดับชาติครั้งที่ 2 กาญจนบุรี : มหาวิทยาลัยราชภัฏกาญจนบุรี.

รัชพล กลัดชื่น, & จรัญ แสนราช. (2561). การเปรียบเทียบประสิทธิภาพอัลกอริทึมและการคัดเลือกคุณลักษณะที่เหมาะสม เพื่อการทำนายผลสัมฤทธิ์ทางการเรียนของนักศึกษาระดับอาชีวศึกษา. วารสารวิจัย มหาวิทยาลัยเทคโนโลยีราชมงคลธัญบุรี, 17(1), 1-10.

ศศิธร ตุ้มเพชรรัตน์, สมบูรณ์ อเนกฤทธิ์มงคล, & สุมนา เกษมสวัสดิ์. (2560). การพยากรณ์ผลการสอบ *TOEIC* สาขาวิชาภาษาอังกฤษ มหาวิทยาลัยรังสิต โดยใช้เทคนิคเหมืองข้อมูล. Paper presented at the การประชุมวิชาการระดับชาติ มหาวิทยาลัยรังสิต ประจำปี 2560.

สมชาย รัตนทองคำ. (2554). เอกสารประกอบการสอน การวัดและประเมินผลทางการศึกษา.

เสกสรรค์ วิลัยลักษณ์, วิภา เจริญภัณฑารักษ์, & ดวงดาว วิชาดากุล. (2558). การใช้เทคนิคการทำเหมืองข้อมูลเพื่อพยากรณ์ผลการเรียนของนักเรียนโรงเรียนสาธิตแห่งมหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตกำแพงแสน ศูนย์วิจัยและพัฒนาการศึกษา. วาสารสาขาวิทยาศาสตร์และเทคโนโลยี, 2(2), 1-17.

อดุลย์ ยิ้มงาม. (2551). การทำเหมืองข้อมูล (Data Mining). Retrieved from สืบค้นเมื่อวันที่ 30 มกราคม 2564, จาก http://compcenter.bu.ac.th/news-information/data-mining

อัจฉราภรณ์ จุฑาผาด. (2559). การพัฒนาระบบสารสนเทศเพื่อการพยากรณ์จำนวนนักศึกษาใหม่โดยใช้กฎการจำแนกต้นไม้ตัดสินใจ. เอกสารประการประชุมวิชาการระดับชาติ. Paper presented at the นเรศวรวิจัย ครั้งที่ 12 : วิจัยและนวัตกรรมกับการพัฒนาประเทศ, พิษณุโลก : มหาวิทยาลัยนเรศวร.

ประวัติผู้เขียน

ชื่อ-สกุล	จิราภรณ์ เจริญยิ่ง
วัน เดือน ปี เกิด	06 มีนาคม 2534
สถานที่เกิด	กรุงเทพฯ
วุฒิการศึกษา	พ.ศ. 2556 ศิลปศาสตร์บัณฑิต เกียรตินิยมอันดับ 2 สาขาบรรณารักษศาสตร์และสารสนเทศศาสตร์ จาก มหาวิทยาลัยศรีนครินทรวิโรฒ