แนวโน้มการเลือกแผนการเรียนมัธยมศึกษาตอนปลาย
ด้วยการใช้ MACHINE LEARNING

ฐิติชัย รักบำรุง

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์และโทรคมนาคม
วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์
มหาวิทยาลัยธุรกิจบัณฑิตย์
ปีการศึกษา 2564

TREND IN THE SELECTION OF A HIGH SCHOOL STUDY PLAN USING MACHINE LEARNING

THITICHAI RUCKBUMRUNG

A Thesis Submitted in Partial Fulfilment of the Requirements

for the Degree of Master of Engineering

Department of Computer and Telecommunication Engineering

College of Innovative Technology And Engineering,

Dhurakij Pundit University

Academic Year 2021

ใบรับรองวิทยานิพนธ์

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์

มหาวิทยาลัยธุรกิจบัณฑิตย์

ปริญญา วิศวกรรมศาสตรมหาบัณฑิต

หัวข้อวิทยานิพนธ์ แนวโน้มการเลือกแผนการเรียนมัธยมศึกษาตอนปลายด้วยการใช้ Machine Learning

เสนอโดย นายฐิติชัย รักบำรุง

สาขาวิชา วิศวกรรมคอมพิวเตอร์และโทรคมนาคม

อาจารย์ที่ปรึกษาวิทยานิพนธ์ อาจารย์ ดร.ชัยพร เขมะภาตะพันธ์

ได้พิจารณาเห็นชอบโดยคณะกรรมการสอบวิทยานิพนธ์แล้ว

.....ประธานกรรมการ

(รองศาสตราจารย์ ดร.ลัญฉกร วุฒิสิทธิกุลกิจ)

.....กรรมการและอาจารย์ที่ปรึกษาวิทยานิพนธ์

(อาจารย์ ดร.ชัยพร เขมะภาตะพันธ์)

.....กรรมการ

(อาจารย์ ดร.ธนัญ จารุวิทยโกวิท)

.....กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.มัชฌิกา อ่องแตง)

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์รับรองแล้ว

.....คณบดีวิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์

(อาจารย์ ดร.ชัยพร เขมะภาตะพันธ์)

วันที่ 23 เดือน พ.ย. พ.ศ. 2564

หัวข้อวิทยานิพนธ์	แนวโน้มการเลือกแผนการเรียนมัธยมศึกษาตอนปลาย ด้วยการใช้ Machine Learning
ชื่อผู้เขียน	ฐิติชัย รักบำรุง
อาจารย์ที่ปรึกษา	ดร.ชัยพร เขมะภาตะพันธ์
สาขาวิชา	วิศวกรรมคอมพิวเตอร์และโทรคมนาคม
ปีการศึกษา	2564

บทคัดย่อ

การเลือกแผนศึกษาต่อเป็นเรื่องสำคัญสำหรับนักเรียน เนื่องจากแผนการเรียนในระดับมัธยมศึกษาตอนปลายจะมีผลต่อการเลือกศึกษาในระดับมหาวิทยาลัย การตัดสินใจเลือกแผนการเรียนเป็นเรื่องยาก เพราะนักเรียนส่วนใหญ่ ยังขาดประสบการณ์และการทำความเข้าใจต่อตนเอง แนวทางหนึ่ง คือครูแนะแนวจะนำข้อมูลผู้เรียนมาวิเคราะห์หาแนวโน้มการศึกษาต่อเพื่อประกอบการตัดสินใจ โดยทั่วไปมักจะดูที่ผลการเรียนในกลุ่มรายวิชาที่มีผลการเรียนดีเป็นหลัก แต่ก็ไม่สามารถสรุปได้ว่านักเรียนจะต้องเลือกเรียนตามที่มีผลการเรียนเป็นที่น่าพอใจเสมอไป เนื่องจากปัจจัยจำนวนมาก ทำให้ไม่สามารถเก็บสถิติตัวแปรได้ครอบคลุม ด้วยเหตุนี้ จึงมีความจำเป็นต้องหากระบวนการหรือเครื่องมือที่น่าเชื่อถือมาช่วยหาแนวโน้มในการตัดสินใจ งานวิจัยนี้ได้เลือกใช้ Machine Learning ซึ่งเป็นระบบที่สามารถเรียนรู้ได้หลายวิธี

วัตถุประสงค์การวิจัยเพื่อ 1) เพื่อสร้างโมเดลการแนะแนวแผนการศึกษาต่อมัธยมศึกษาตอนปลาย 2) เพื่อเปรียบเทียบประสิทธิภาพของโมเดลการแนะแนวแผนการศึกษาต่อมัธยมศึกษาตอนปลาย ข้อมูลเพื่อใช้ในการสร้าง Machine Learning 258 ชุด ผ่านกระบวนการ 3 แบบ ข้อมูลชุดที่ 1 กระบวนการ Correlation Analysis ข้อมูลชุดที่ 2 ใช้วิธีการลดขนาดข้อมูล (PCA) ข้อมูลชุดที่ 3 Feature Selector (Sequential Feature Selector : Backward Selection) โดยการแบ่งข้อมูลออกเป็น 2 กลุ่ม คือ ชุดเรียนรู้ (Training Set) ร้อยละ 80 และข้อมูลชุดทดสอบ (Testing Set) ร้อยละ 20 ผลการวิจัยพบว่า ค่าความถูกต้อง (Accuracy) ของข้อมูลชุดที่ 1 – 3 ตามลำดับดังนี้ 1) Decision Tree ร้อยละ 79,79,87 2) Artificial Neural Network ร้อยละ 83, 85, 83 3) Naive Bayes ร้อยละ 87, 83, 88 4) Support Vector Machine ร้อยละ 87, 83, 85

คำสำคัญ: Machine Learning, Decision Tree, Naive Bayes, Neural network, Support Vector Machine

Thesis Title	TRENDS IN THE SELECTION OF A HIGH SCHOOL STUDY PLAN USING MACHINE LEARNING
Author	Thitichai Ruckbumrung
Thesis Advisor	Dr. Chaiyaporn Khemapatapan
Department	Computer and Telecommunications Engineering
Academic Year	2021

ABSTRACT

Selecting a high school study plan is deemed important for junior high school students because such plan can have a profound impact on the choice to opt into their university studies. This can be a very tough decision for most students due to their lack of experience and self-understanding. To help their students select a study plan, one solution is for guidance teachers to use their students' data to analyze the trends in their further studies in order to support decision making. In general, the learning outcomes of the subject group which is satisfactory are mainly considered. However, it cannot be concluded that students will always make a selection according to a study plan with satisfying grade results. Due to a number of factors, it is impossible to collect all comprehensive variables. It is thus necessary to find a reliable process or tool to help identify trends in decision making. In this research, Machine Learning, a system that can learn in many ways, was utilized.

The objective of this research was twofold: 1) to create a guidance model for high school study plans, and 2) to compare its efficiency. There were 258 data sets for creating the Machine Learning through three processes; the data set 1 was Correlation Analysis; the data set 2 was the Data Reduction (PCA); and the data set 3 was Feature Selector (Sequential Feature Selector:Backward Selection). The data was then divided into two groups: the training set accounting for 80% and the testing set representing 20%. The results showed the accuracy of data sets 1 to 3 in order as follows: 1) Decision Tree, which was 79%, 79% , and 87%, 2) Artificial Neural Network, which was 83%, 85%, and 83%, 3) Naive Bayes, which was 87%, 83%, and 88%, and 4) Support Vector Machine, which was 87%, 83%, and 85%.

Keywords : Machine Learning, Decision Tree, Naive Bayes, Neural Network, Support Vector Machine

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จได้ด้วยความกรุณาจาก อาจารย์ ดร.ชัยพร เขมะภาตะพันธ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ได้กรุณาให้คำแนะนำแก้ไขข้อบกพร่อง และความช่วยเหลือในการทำวิจัยตลอดระยะเวลาการศึกษา ขอขอบพระคุณเป็นอย่างสูง

ผู้วิจัยขอขอบพระคุณ รองศาสตราจารย์ ดร.ลัญฉกร วุฒิสิทธิกุลกิจ ที่กรุณาให้เกียรติเป็นประธานสอบวิทยานิพนธ์ อีกทั้งผู้ช่วยศาตราจารย์ ดร.มัชฌิกา อ่องแตง และอาจารย์ ดร.ธนัญ จารุวิทยโกวิท ที่กรุณาเป็นกรรมการในการสอบวิทยานิพนธ์ ที่ให้ข้อเสนอแนะในกระบวนการวิจัย ทำให้วิทยานิพนธ์ฉบับนี้สมบูรณ์มากยิ่งขึ้น

ขอขอบคุณ ดร.ศรัณย์ ภิบาลชนม์ ที่กรุณาให้คำแนะนำด้านเทคนิค Machine Learning ทำให้ผู้วิจัยจัดการเครื่องมือได้อย่างมีประสิทธิภาพ

ขอขอบคุณ นางสาวกุลธรา อานนท์ และเจ้าหน้าที่มหาวิทยาลัยธุรกิจบัณฑิตย์ทุกท่าน ที่ให้ความสะดวกในการประสานงานของผู้วิจัยด้วยดีเสมอมา

ท้ายนี้ ขอกราบขอบพระคุณบิดา มารดา และครอบครัว ตลอดจนครู อาจารย์ ที่ให้ความเมตตา ชี้แนะ เป็นกำลังใจแก่ผู้วิจัย จึงขอกราบขอบพระคุณผู้มีส่วนเกี่ยวข้องกับการทำวิทยานิพนธ์เป็นอย่างสูงไว้ ณ โอกาสนี้

ฐิติชัย รักบำรุง

สารบัญ

หน้า

บทคัดย่อภาษาไทย..... ฆ
บทคัดย่อภาษาอังกฤษ..... จ
กิตติกรรมประกาศ..... ฉ
สารบัญตาราง..... ซ
สารบัญภาพ..... ฌ
บทที่
1 บทนำ 1
1.1 ความเป็นมาและความสำคัญของปัญหา..... 1
1.2 วัตถุประสงค์การวิจัย..... 3
1.3 ขอบเขตของการวิจัย..... 3
1.4 เครื่องมือในการวิจัย..... 6
1.5 ประโยชน์ที่คาดว่าจะได้รับ..... 6
1.6 นิยามศัพท์เฉพาะ..... 6
1.7 แผนการดำเนินงานวิจัย..... 7
1.8 การตอบรับการตีพิมพ์..... 8
1.9 โครงสร้างของรายงานวิจัยส่วนที่เหลือ..... 8
2 เอกสารและงานวิจัยที่เกี่ยวข้อง..... 9
2.1 แนวคิดการแนะแนว..... 9
2.2 การเรียนรู้ของเครื่อง (Machine Learning)..... 17
2.3 ขั้นตอนวิธี (Algorithms)..... 24
2.4 การวิเคราะห์ตัวเลือกคุณลักษณะตามลำดับ (Feature Selector: Backward Selection)..... 32
2.5 การวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis : PCA)..... 32
2.6 ซอฟต์แวร์และโปรแกรมที่ใช้ในการศึกษา..... 33
2.7 เอกสารงานวิจัยที่เกี่ยวข้อง..... 39

สารบัญ (ต่อ)

บทที่	หน้า
3 การดำเนินงานวิจัย	41
3.1 การเลือกข้อมูล (Selection)	41
3.2 การกลั่นกรองข้อมูล (Data cleansing)	43
3.3 การแปลงข้อมูล (Data Transformation)	43
3.4 การสร้างแบบจำลอง (Model)	54
3.5 การแปรผลและการประเมินผล (Interpretation Evaluation)	57
4 ผลการวิจัย	59
5 สรุปและอภิปรายผล	78
5.1 อภิปรายผล	80
5.2 ปัญหาและข้อเสนอแนะ	81
บรรณานุกรม	82
ภาคผนวก	87
ก ภาพต้นไม้ตัดสินใจด้วยวิธีการ Correlation Analysis	88
ข ภาพต้นไม้ตัดสินใจด้วยวิธีการ Principal Analysis : PCA	90
ค ภาพต้นไม้ตัดสินใจด้วยวิธีการ Feature Selection : Backward selection	92
ประวัติผู้เขียน	94

สารบัญตาราง

ตารางที่ **หน้า**

1.1 แผนดำเนินงาน..... 7

3.1 การกำหนดตัวแปร..... 42

3.2 การรวมเอกสารอิเล็กทรอนิกส์ในไฟล์ Excel…….. 43

3.3 การแปลงข้อมูลและการกำหนด Code……. 44

3.4 แสดงตัวอย่างข้อมูลและ Code ของ Attributes แบบ Clustering…….. 47

3.5 แสดงตัวอย่างรูปแบบข้อมูลและ Code ข้อมูล..... 47

4.1 ผลการทดสอบแบบจำลองชุดที่ 1 ด้วยเทคนิค Decision Tree……. 60

4.2 ผลการทดสอบแบบจำลองข้อมูลชุดที่ 1 ด้วยเทคนิค Artificial Neural Network. 61

4.3 ผลการทดสอบแบบจำลองข้อมูลชุดที่ 1 ด้วยเทคนิค Naive Bayes…….. 62

4.4 ผลการทดสอบแบบจำลองข้อมูลชุดที่ 1 ด้วยเทคนิค Support Vector Machine... 63

4.5 ผลการทดสอบแบบจำลองชุดที่ 2 ด้วยเทคนิค Decision Tree……. 63

4.6 ผลการทดสอบแบบจำลองชุดที่ 2 ด้วยเทคนิค Artificial Neural Network…….. 64

4.7 ผลการทดสอบแบบจำลองข้อมูลชุดที่ 2 ด้วยเทคนิค Naive Bayes…….. 65

4.8 ผลการทดสอบแบบจำลองข้อมูลชุดที่ 2 ด้วยเทคนิค Support Vector Machine... 66

4.9 ผลการทดสอบแบบจำลองชุดที่ 3 ด้วยเทคนิค Decision Tree……. 66

4.10 ผลการทดสอบแบบจำลองชุดที่ 3 ด้วยเทคนิค Artificial Neural Network…… 68

4.11 ผลการทดสอบแบบจำลองชุดที่ 3 ด้วยเทคนิค Naive Bayes…….. 69

4.12 ผลการทดสอบแบบจำลองข้อมูลชุดที่ 3 ด้วยเทคนิค Support Vector Machine. 69

4.13 การเปรียบเทียบเทคนิค Decision Tree ด้วยข้อมูลชุดที่ 1 ชุดที่ 2 และชุดที่ 3.... 70

4.14 การเปรียบเทียบเทคนิค Artificial Neural Network ด้วยข้อมูลชุดที่ 1 ชุดที่ 2 และชุดที่ 3..... 72

4.15 การเปรียบเทียบเทคนิค Naive Bayes ด้วยข้อมูลชุดที่ 1 ชุดที่ 2 และชุดที่ 3..... 73

4.16 การเปรียบเทียบเทคนิค Support Vector Machine ด้วยข้อมูลชุดที่ 1 ชุดที่ 2 และชุดที่ 3..... 75

4.17 สรุปผลการวิเคราะห์ค่า Accuracy ด้วยข้อมูลชุดที่ 1 ชุดที่ 2 และชุดที่ 3..... 76

สารบัญภาพ

ภาพที่ **หน้า**

2.1 ระบบ AI..... 17

2.2 คำสั่งตามตรรกะแบบเดิม (Logical statement)..... 18

2.3 ประเภทการเรียนรู้ของ Machine Learning..... 19

2.4 การเรียนรู้แบบมีผู้ดูแล (Supervised learning)..... 19

2.5 การจำแนก Classification..... 20

2.6 การถดถอยเชิงเส้น..... 21

2.7 Scatter matrix..... 23

2.8 การจำแนกคุณลักษณะของลูกค้าที่ทำการซื้อคอมพิวเตอร์..... 25

2.9 กฎการตัดสินใจแบบ if-then-else..... 26

2.10 Decision Tree Classifier..... 27

2.11 การทำนายคลาส..... 28

2.12 การจัดประเภทแบบหลายคลาส..... 28

2.13 การพล๊อต Tree ด้วยฟังก์ชั่น plot_tree: tree.plot_tree(clf)..... 28

2.14 โครงข่ายประสาท (Neural Network)..... 30

2.15 การวิเคราะห์ของทฤษฎีเบย์..... 31

2.16 เส้นแบ่งกลุ่มข้อมูล Hyper plane (SVM)..... 31

2.17 ตัวอย่างกระบวนการทำ PCA..... 33

2.18 แสดงรูปแบบของการรองรับการติดตั้งที่หลากหลายของระบบปฏิบัติการทั้ง Window, MacOS, Linux..... 34

2.19 มาตรฐานการประมวลผลเชิงโต้ตอบในภาษาโปรแกรม..... 35

3.1 การเลือกข้อมูลเพื่อใช้ในการสร้าง Machine Learning..... 48

3.2 การวิเคราะห์ความสัมพันธ์ระหว่าง Attributes ด้วยการสร้าง Scatter Plot..... 49

3.3 กราฟแสดงความสัมพันธ์ข้อมูลกลุ่มภูมิลำเนา..... 49

3.4 กราฟแสดงความสัมพันธ์ข้อมูลกลุ่มอาชีพของผู้ปกครอง..... 50

3.5 กราฟแสดงความสัมพันธ์ข้อมูลกลุ่มรายได้ของผู้ปกครอง..... 50

3.6 การวิเคราะห์ความสัมพันธ์ด้วยการสร้าง Scatter Plot..... 51

สารบัญภาพ (ต่อ)

ภาพที่ **หน้า**

3.7 การเริ่มต้น PCA และการทำสเกลข้อมูล……..... 52

3.8 ผลของ PCA จากค่า Variance……..... 53

3.9 การใช้คำสั่ง Feature Selector: Backward Selection…… 53

3.10 การ Download และการติดตั้ง Anaconda…..……. 54

3.11 Integrated Development Environment……..… 55

3.12 การใช้ Jupyter notebook ในการทำ Machine Learning..... 55

3.13 การสร้างโมเดล Decision Tree…….. 56

3.14 การสร้างโมเดล Artificial Neural Network……..... 56

3.15 การสร้างโมเดล Naive Bayes…… 57

3.16 การสร้างโมเดล Support Vector…… 57

3.17 แผนผังขั้นตอนการวิจัย..... 58

4.1 แผนผังต้นไม้จากการทำนายข้อมูลชุดที่ 1 ด้วยเทคนิค Decision Tree..... 60

4.2 แผนผังต้นไม้จากการทำนายข้อมูลชุดที่ 3 ด้วยเทคนิค Decision Tree……. 67

5.1 การเปรียบเทียบค่าความถูกต้อง(Accuracy)ด้วยข้อมูลชุดที่ 1 ชุดที่ 2 และชุดที่ 3 79

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

เมื่อศึกษาอยู่ชั้นมัธยมศึกษาปีที่ 3 นักเรียนทุกคนต้องตัดสินใจเลือกแผนการเรียนต่อมัธยมศึกษาตอนปลาย ระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์หรือแผนศิลป์ การเลือกแผนศึกษาต่อเป็นเรื่องสำคัญสำหรับนักเรียนมัธยมศึกษาตอนต้น เนื่องจากแผนการเรียนในระดับมัธยมศึกษาตอนปลายจะมีผลต่อการเลือกเรียนในระดับมหาวิทยาลัยต่อไป ซึ่งเป็นตัวกำหนดด้านการประกอบอาชีพในอนาคตของนักเรียน อย่างไรก็ตาม การตัดสินใจเลือกแผนการศึกษาต่อในระดับมัธยมศึกษาตอนปลายเป็นเรื่องยากของนักเรียนส่วนใหญ่ เพราะนักเรียนยังขาดประสบการณ์และการทำความเข้าใจต่อตนเอง การได้คำแนะนำส่วนใหญ่จะมาจากคนในครอบครัว ซึ่งแต่ละครอบครัวก็อาจจะมีแนวคิดที่แตกต่างกัน ยังไม่รวมถึงปัจจัยอื่นๆ อาทิ ความชอบหรือความถนัดส่วนตัวของนักเรียน ผลการเรียนของนักเรียน เป็นต้น ปัจจัยเหล่านี้ทำให้นักเรียนเกิดความไม่แน่ใจหรืออาจเลือกแผนการเรียนไม่เหมาะสมตามศักยภาพ แนวทางหนึ่งในการเลือกแผนการเรียนมัธยมศึกษาตอนปลาย คือการนำข้อมูลผู้เรียนมาวิเคราะห์หาแนวโน้มการศึกษาต่อเพื่อประกอบการตัดสินใจ ข้อมูลผลการเรียนเป็นสิ่งหนึ่งที่ครูแนะแนวจะนำมาใช้เป็นข้อมูลเชิงประจักษ์ที่แสดงศักยภาพของผู้เรียนเป็นตัวบอกแนวโน้มที่จะนำมาวิเคราะห์ ข้อสังเกตคือ เมื่อเริ่มพิจารณาแนวทางการเลือกแผนการเรียนมักจะดูที่ผลการเรียนในกลุ่มรายวิชาใดวิชาหนึ่งที่ได้ผลการเรียนดีเป็นหลัก หากเรียนวิชาทางด้านคณิตศาสตร์และวิทยาศาสตร์ได้ดีหรือเรียนวิชาทางด้านสังคมและภาษาได้เด่น ก็มีแนวโน้มว่าจะแนะนำหรือตัดสินใจเลือกเรียนทางด้านสาขาวิชาที่ผลการเรียนเป็นที่น่าพอใจนั้น แต่ข้อสรุปนี้ก็ไม่ได้ตัดสินหรือสรุปได้ว่านักเรียนที่มีผลการเรียนคณิตศาสตร์และวิทยาศาสตร์หรือสังคมและภาษาดีนั้น จะต้องเลือกเรียนตามแผนการเรียนที่มีผลน่าพอใจนั้นๆ ครูแนะแนวไม่สามารถจะตัดสินใจได้อย่างแน่ชัด และมีปัจจัยจำนวนมากที่อาจส่งผลต่อการตัดสินใจ แต่ก็ไม่สามารถเก็บสถิติตัวแปรได้ครอบคลุมตั้งแต่การเริ่มต้นการเรียนมาวิเคราะห์ผลทั้งหมดได้ด้วยวิธีการคิดการตัดสินใจของตัวนักเรียนหรือครูแนะแนวได้ด้วยตนเอง

ทำให้จำเป็นต้องหากระบวนการหรือเครื่องมือที่น่าเชื่อถือมาช่วยหาแนวโน้มในการตัวสินใจตัวเลือกหนึ่งที่น่าสนใจก็คือ Machine Learning เป็นระบบที่สามารถเรียนรู้ได้หลายวิธีและวิธีหนึ่งที่นำเอาข้อมูลที่มีมาใช้คือ การเรียนรู้จากต้นแบบด้วยตนเองเป็นเครื่องมือช่วยประมวลผลข้อมูล สามารถสร้างแบบจำลองมีความแม่นยำในการทำนายสูง ประยุกต์ใช้ได้ในหลากหลายงานเป็นที่ยอมรับระดับสากล ดังตัวอย่างงานวิจัยการใช้ Machine Learning สำหรับการแนะนำการเรียนของนักศึกษา ได้นำวิธีการนี้มาใช้ในการวิเคราะห์การเรียนที่มีความเหมาะสมกับนักศึกษา การเรียนรู้ของเครื่องมีหลากหลายตัวเลือก เช่น Rapidminer, Tensorflow, Weka, Scikit-learn เป็นต้น การเลือกใช้งาน Package หรือ Program เหล่านี้ขึ้นอยู่กับลักษณะงานของผู้ใช้ จำนวนข้อมูล ความถนัดในการเขียนโปรแกรมหรือการใช้งานคำสั่งหลัก ในงานวิจัยนี้เลือก Scikit-learn ซึ่งจะมีข้อจำกัดน้อยกว่าในการใช้งานตัวเลือกอื่นเพราะเป็นโอเพนซอร์ซ (Open Source) ถูกใช้บนภาษา Python มีการทำงานแบบการแปลชุดคำสั่งทีละบรรทัดทำงานได้อิสระต่างจาก Package หรือ Program อื่นที่มีข้อจำกัดการใช้งานมากกว่า

งานวิจัยนี้เลือกใช้การเรียนรู้ของเครื่องในแบบ Supervised Learning ประกอบด้วย 4 วิธี คือ 1) Decision Tree 2) Artificial Neural Network 3) Naive Bayes 4) Support Vector Machine เนื่องจาก 1) Decision Tree เป็นวิธีการที่ง่ายที่สุดที่เราจะสามารถตรวจสอบข้อมูลโดยเป็นวิธีที่สามารถแสดงผลลัพธ์แสดงความสัมพันธ์ออกมาเป็นรูปแบบตารางต้นไม้ที่ประกอบด้วยกิ่งและก้านโดยจะแทนลักษณะประจำที่นำมาใช้แยกกลุ่มข้อมูล ใบจะแทนค่าของคลาสหรือการกระจายของคลาส 2) Artificial Neural Network เป็นวิธีที่สามารถหาความสัมพันธ์เมื่อมีอินพุตโดเมนหรือฟังก์ชันและได้ผลลัพธ์ออกมา โครงข่ายประสาทเทียมโดยปกติจะเป็นเลเยอร์โหนดของกราฟสอดคล้องกับเซลล์ประสาทและการเชื่อมโยงไปยังซินแนปส์ Artificial Neural Network จะทำให้เราทราบว่าฟังก์ชันที่เหมาะสมกับโดเมนหรือ input ว่าคืออะไร 3) Naive Bayes เป็นวิธีที่ใช้วิเคราะห์ความสัมพันธ์ระหว่างตัวแปรในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์เหมาะกับกรณีของเซตตัวอย่างจำนวนมากและคุณสมบัติ (Attributes) ของตัวอย่างไม่ขึ้นต่อกันโดยการกำหนดให้ความน่าจะเป็นของข้อมูลเท่ากับสมการ 4) Support Vector Machine เป็นวิธีที่ได้รับความนิยมในการทำข้อมูลลักษณะแยกประเภทข้อมูลได้ดี

1.2 วัตถุประสงค์

1. เพื่อสร้างโมเดลการแนะแนวแผนการศึกษาต่อมัธยมศึกษาตอนปลาย

2. เพื่อเปรียบเทียบประสิทธิภาพของโมเดลการแนะแนวแผนการศึกษา ต่อมัธยมศึกษาตอนปลายระหว่างวิธีการ Decision Tree, Artificial Neural Network, Naive Bayes, Support Vector Machine

1.3 ขอบเขตการวิจัย

1.3.1 ขอบเขตด้านข้อมูล

ข้อมูลที่ใช้ในงานวิจัยนี้เป็น Attributes ข้อมูลของผู้สำเร็จชั้นมัธยมศึกษาปีที่ 4 - 6 ปีการศึกษา 2554 - 2556 จำนวน 454 คน ประกอบด้วย 70 Attributes ดังต่อไปนี้

1) เพศ

2) ภูมิลำเนา

3) อาชีพของผู้ปกครอง

4) รายได้ผู้ปกครอง

5) แผนการเรียน

6) วิชาคณิตศาสตร์

ภาคการศึกษาที่ 1-1

ภาคการศึกษาที่ 2-1

ภาคการศึกษาที่ 3-1, 3-2

ภาคการศึกษาที่ 4-1, 4-2

ภาคการศึกษาที่ 5-1, 5-2

ภาคการศึกษาที่ 6-1

7) วิชาวิทยาศาสตร์

ภาคการศึกษาที่ 1-1

ภาคการศึกษาที่ 2-1

ภาคการศึกษาที่ 3-1

ภาคการศึกษาที่ 4-1

ภาคการศึกษาที่ 5-1

ภาคการศึกษาที่ 6-1

8) วิชาภาษาอังกฤษ

ภาคการศึกษาที่ 1-1
ภาคการศึกษาที่ 2-1
ภาคการศึกษาที่ 3-1
ภาคการศึกษาที่ 4-1
ภาคการศึกษาที่ 5-1
ภาคการศึกษาที่ 6-1

9) วิชาภาษาไทย

ภาคการศึกษาที่ 1-1
ภาคการศึกษาที่ 2-1
ภาคการศึกษาที่ 3-1
ภาคการศึกษาที่ 4-1
ภาคการศึกษาที่ 5-1
ภาคการศึกษาที่ 6-1

10) วิชาสังคม

ภาคการศึกษาที่ 1-1
ภาคการศึกษาที่ 2-1
ภาคการศึกษาที่ 3-1
ภาคการศึกษาที่ 4-1
ภาคการศึกษาที่ 5-1
ภาคการศึกษาที่ 6-1

11) วิชาศิลปะ

ภาคการศึกษาที่ 1-1
ภาคการศึกษาที่ 2-1
ภาคการศึกษาที่ 3-1
ภาคการศึกษาที่ 4-1
ภาคการศึกษาที่ 5-1
ภาคการศึกษาที่ 6-1

12) วิชาสุขศึกษา

ภาคการศึกษาที่ 1-1

ภาคการศึกษาที่ 2-1

ภาคการศึกษาที่ 3-1

ภาคการศึกษาที่ 4-1

ภาคการศึกษาที่ 5-1

ภาคการศึกษาที่ 6-1

13) วิชาประวัติศาสตร์

ภาคการศึกษาที่ 1-1

ภาคการศึกษาที่ 2-1

ภาคการศึกษาที่ 3-1

ภาคการศึกษาที่ 4-1

ภาคการศึกษาที่ 5-1

ภาคการศึกษาที่ 6-1

14) วิชาพลศึกษา

ภาคการศึกษาที่ 1-1

ภาคการศึกษาที่ 2-1

ภาคการศึกษาที่ 3-1

ภาคการศึกษาที่ 4-1

ภาคการศึกษาที่ 5-1

ภาคการศึกษาที่ 6-1

15) วิชาการงานอาชีพและเทคโนโลยี

ภาคการศึกษาที่ 1-1, 2-1, 3-1, 4-1, 5-1, 6-1

ภาคการศึกษาที่ 2-1

ภาคการศึกษาที่ 3-1

ภาคการศึกษาที่ 4-1

ภาคการศึกษาที่ 5-1

ภาคการศึกษาที่ 6-1

16) เกรดเฉลี่ยมัธยมศึกษาปีที่ 3

17) เกรดเฉลี่ยมัธยมศึกษาปีที่ 6

1.3.2 ขอบเขตด้านอัลกอริทึมที่นำมาใช้ในการสร้างโมเดลในการทำนายแผนการศึกษาต่อมัธยมศึกษาตอนปลาย

1) Decision Tree

2) Naive Bayes

3) Artificial Neural Network

4) Support Vector Machine

1.4 เครื่องมือในการวิจัย

1.4.1 ด้านซอฟต์แวร์และโปรแกรม (Software and Program) มีดังนี้

1) Anaconda

2) Scikit - Learn

3) Numpy

4) Pandas

5) Matplotlib

6) Jupyter Notebook (IDE : Integrated Development Environment)

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1.5.1 ครูแนะแนวการศึกษา ได้ข้อมูลช่วยในการตัดสินใจแนะนำการเลือกแผนการศึกษาต่อของผู้เรียนชั้นมัธยมศึกษาตอนปลาย

1.5.2 ได้โมเดลการหาแนวโน้มการเลือกแผนการศึกษาต่อมัธยมศึกษาตอนปลายที่มีประสิทธิภาพ ทำให้กระบวนการแนะแนวของครูมีเครื่องมือที่มีประสิทธิภาพดีขึ้น

1.6 นิยามศัพท์เฉพาะ

1.6.1 การแนะแนว หมายถึง กระบวนการแนะแนวของครูที่ใช้ Machine Learning เป็นส่วนหนึ่งช่วยแนะนำผู้เรียนในการตัดสินใจเลือกแผนการเรียนต่อมัธยมศึกษาตอนปลาย ระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์หรือแผนศิลป์

1.6.2 Scikit – learn หมายถึง Bundleที่รวบรวม Library แบบโอเพนซอร์ซ สำหรับการพัฒนาโปรแกรมโดยใช้ Machine Learning ในงานวิจัยนี้ ใช้ภาษา Python ทำฟังก์ชันในการแบ่งประเภท

ข้อมูล การแบ่งกลุ่มข้อมูล การวิเคราะห์ข้อมูลและสร้างโมเดลเพื่อหาแนวโน้มการเลือกแผนการศึกษาต่อมัธยมศึกษาตอนปลาย

1.6.3 Machine Learning หมายถึง กระบวนการเรียนรู้ด้วยตนเองของเครื่องจากตัวอย่าง โดยไม่ได้ต้องป้อนคำสั่งกำกับตลอดทุกขั้นตอน แต่ได้ผลลัพธ์ที่แม่นยำ สร้างอัลกอริทึมที่สามารถเรียนรู้ข้อมูล และทำนายข้อมูลได้ ในงานวิจัยนี้ผู้วิจัยใช้วิธี Decision Tree, Naive Bayes, Artificial Neural Network และ Support Vector Machine

1.6.4 นักเรียนที่ประสบผลสำเร็จในการเรียน หมายถึง นักเรียนที่ศึกษาในสายที่ตนเลือกเรียนต่อแล้วมีผลการเรียนเฉลี่ยเมื่อสำเร็จการศึกษาอยู่ที่ระดับ 3.00 ขึ้นไป ใช้ข้อมูลนี้ในการสร้าง Machine Learning

1.7 แผนการดำเนินงาน

ในการดำเนินการวิจัยได้กำหนดแผนดำเนินงาน ดังตารางที่ 1

ตารางที่ 1.1 แผนดำเนินงาน

รายการดำเนินงาน	**ระยะเวลา (เดือน)**											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1.ศึกษาเอกสาร ตำรา งานวิจัยที่เกี่ยวข้อง												
2.ทำ Data Preparation เพื่อคัดเลือกข้อมูลให้มีความพร้อมสำหรับการวิเคราะห์												
3.ทำ Modeling โดยนำข้อมูลเข้ามาวิเคราะห์เพื่อสร้างแบบจำลองเพื่อหากฎความสัมพันธ์												
4.วัดประสิทธิภาพของผลลัพธ์ (Evaluation)												
5.วิเคราะห์และสรุปผลการทดลอง												
6.จัดทำรูปเล่มฉบับสมบูรณ์												

1.8 การตอบรับการตีพิมพ์

หัวข้อ “แนวโน้มการเลือกแผนการเรียนมัธยมศึกษาตอนปลาย โดยใช้การเรียนรู้ของเครื่องร่วมกับ SCIKIT-LEARN PACKAGE”

TREND OF SELECTION FOR HIGH SCHOOL STUDY PLANING USING MACHINE LEARNING นำเสนอ ณ การประชุมวิชาการระดับชาติวิทยาศาสตร์และเทคโนโลยี ครั้งที่ 5 (The 5th National Conference on Science and Technology) 2021 เมื่อวันที่ 15 ม.ค. 2564

1.9 โครงสร้างของรายงานวิจัยส่วนที่เหลือ

บทที่ 2 แนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้อง

บทที่ 3 การดำเนินงานวิจัย

บทที่ 4 การอภิปรายผล

บทที่ 5 บทสรุปและข้อเสนอแนะ

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

แนวโน้มการเลือกแผนการเรียนมัธยมศึกษาตอนปลาย โดยใช้การเรียนรู้ของเครื่องร่วมกับ SCIKIT-LEARN PACKAGE ผู้วิจัยศึกษาเอกสารงานวิจัยที่เกี่ยวข้องในประเด็นต่อไปนี้

2.1 แนวคิดการแนะแนว

2.2 การเรียนรู้ของเครื่อง (Machine Learning)

2.3 ขั้นตอนวิธี (Algorithms)

2.4 การวิเคราะห์ตัวเลือกคุณลักษณะตามลำดับ(Feature Selector : Backward Selection)

2.5 การวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis : PCA)

2.6 ซอฟต์แวร์และโปรแกรมที่ใช้ในการศึกษา

2.7 เอกสารงานวิจัยที่เกี่ยวข้อง

2.1 แนวคิดการแนะแนว

2.1.1 ความหมายของการแนะแนว

พจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ. 2554 ได้บัญญัติศัพท์คำว่า “แนะแนว” ไว้ว่าหมายถึง “แนะนำแนวทางว่าควรปฏิบัติอย่างไร” ดังนั้นในศาสตร์ทางการแนะแนว จึงมีนักการศึกษาทางด้านการแนะแนวได้ให้ความหมายไว้มากมาย เช่น กรมวิชาการ กล่าวว่า “การแนะแนวคือจิตวิทยาประยุกต์แขนงหนึ่งที่ว่าด้วยการสร้างและพัฒนาคนให้รู้จักช่วยเหลือตนเองหรือพึ่งตนเองได้โดยกระบวนการที่ส่งเสริมให้บุคคลได้มีบทบาทเต็มที่ในการเรียนรู้เพื่อที่จะพัฒนาศักยภาพ และสามารถจัดการกับชีวิตของตนอย่างฉลาด” ทั้งนี้ หากกล่าวถึงในมิติทางการศึกษา นงลักษณ์ ประเสริฐและจรินทร วินทะวิไชย์ ได้นิยามคำว่าการแนะแนวไว้ว่า “กระบวนการต่อเนื่องที่ใช้ทั้งศาสตร์และศิลป์ เพื่อช่วยเหลือบุคคลให้รู้จักตนเองและสิ่งแวดล้อม สามารถตัดสินใจเลือกสิ่งต่างๆ ได้ด้วยตนเองอย่างฉลาด มีเหตุผล รู้จักป้องกันปัญหา วางแผน และพัฒนาตนเองได้เต็มศักยภาพ” และแนวคิดของ อัชรา เอิบสุขสิริ กล่าวว่า “การแนะแนว หมายถึง กระบวนการช่วยเหลือบุคคลให้รู้จักและเข้าใจตนเอง เข้าใจสภาพแวดล้อม สามารถตัดสินใจในการแก้ปัญหาต่างๆ และวางแผนชีวิตได้อย่างฉลาด โดยพัฒนาตนเองให้เติบโตเต็มศักยภาพและดำเนินชีวิตอย่างมีความสุข

และมีคุณค่าต่อสังคม” ทั้งยังสอดคล้องกับสำนักวิชาการและมาตรฐานการศึกษา ซึ่งได้กล่าวถึงความหมายของการแนะแนวว่า “การแนะแนว หมายถึง กระบวนการที่ช่วยให้ผู้เรียนได้รู้จัก เข้าใจ รัก และเห็นคุณค่าในตัวเองและผู้อื่น คิดเป็น ใช้ชีวิตเป็น สามารถตัดสินใจและวางแผนการศึกษา อาชีพ และปรับตัวอยู่ในสังคมได้อย่างมีความสุข” และมิลเลอร์ กล่าวว่า “การแนะแนว เป็นส่วนหนึ่งของกระบวนการทางการศึกษา ซึ่งเกี่ยวข้องกับการช่วยเหลือแต่ละบุคคลให้สามารถเข้าใจตนเอง ตัดสินใจได้ด้วยตนเองและมีการวางแผนในการ พัฒนาระบบของชีวิตของตนเอง”

จากความหมายข้างต้น สามารถกล่าวโดยสรุปได้ว่า “การแนะแนว หมายถึง กระบวนการที่ช่วยเหลือให้บุคคลได้รู้จักตนเอง เข้าใจตนเองและสภาพแวดล้อมและสามารถตัดสินใจเลือกแนวทางการดำเนินชีวิตสำหรับตนเองในด้านการศึกษา ด้านอาชีพและด้านส่วนตัวและสังคมได้อย่างเหมาะสม สามารถป้องกันปัญหา แก้ไขปัญหา และพัฒนาตนเองได้อย่างมีประสิทธิภาพและดำรงชีวิตได้อย่างมีความสุข”

2.1.2 ปรัชญา และจุดมุ่งหมายของการแนะแนว

พนม ลิ้มอารีย์ ได้สรุปหลักการที่สำคัญของการแนะแนว (basic principles of guidance) ไว้ดังนี้

1. การจัดบริการแนะแนวในโรงเรียนจะต้องมุ่งให้ความช่วยเหลือผู้เรียนทุกคน เนื่องจาก ผู้เรียนทุกคนย่อมต้องการความช่วยเหลือจากโรงเรียนของตน และเป็นการให้บริการด้วยความ เสมอภาคเป็นธรรม และเท่าเทียมกัน

2. การจัดบริการแนะแนวจะต้องกระทำอย่างเป็นกระบวนการที่ต่อเนื่อง คือ จัดอย่างมีระบบ มีระเบียบแบบแผน มีความสัมพันธ์ต่อเนื่องกันไปเป็นลูกโซ่ทุกขั้นตอน จนกระทั่งบุคคลที่ได้รับ ความช่วยเหลือสามารถนำตนเองได้ ช่วยตนเองได้

3. ผู้ทำงานแนะแนวจะต้องยอมรับในความเป็นเอกัตบุคคล (individual) ของผู้เรียน นั่นคือ จะต้องมีความเข้าใจและยอมรับในเรื่องความแตกต่างระหว่างบุคคล ซึ่งมีประเด็นสำคัญดังนี้

- บุคคลแต่ละคนย่อมมีลักษณะเฉพาะของตนเองจะไม่เหมือนคนอื่น ไม่ว่ารูปร่าง สติปัญญา ความสามารถ อุปนิสัย ค่านิยม ความสนใจ เป็นต้น

- บุคคลแต่ละคนย่อมมีพัฒนาการไปตามลักษณะเฉพาะของตน อย่างมีลำดับขั้นและต่อเนื่อง

- บุคคลแต่ละคนย่อมมีกระบวนการแห่งการเปลี่ยนแปลงของตน ตามประสบการณ์ที่ตนเองประสบมาและตามแนวทางหรือแผนการของตนที่วางไว้สำหรับอนาคต

4. การแนะแนวเป็นงานที่วางอยู่บนพื้นฐานกระบวนการพฤติกรรมของบุคคล และเกี่ยวข้องกับพัฒนาการของมนุษย์ ดังนั้นการแนะแนวจึงจำเป็นต้องใช้เครื่องมือและกลวิธีต่างๆ ทั้งที่เป็นแบบทดสอบและไม่ใช่แบบทดสอบ เพื่อจะได้เข้าใจบุคคลแต่ละคน และเพื่อช่วยให้บุคคลแต่ละคนได้ เข้าใจตนเอง เพื่อจะได้สามารถควบคุมพัฒนาการส่วนตัวของผู้เรียนได้

5. ผู้ทำงานด้านการแนะแนวจะต้องเคารพในสิทธิและเสรีภาพของบุคคลแต่ละคน นั่นคือ จะต้องยอมรับว่าผู้เรียนแต่ละคนมีอิสรภาพที่จะเลือกแนวทางชีวิตของตนเอง การเลือกและการตัดสินใจของผู้เรียนควรเกิดจากการใช้วิจารณญาณของผู้เรียนเองไม่ใช่เกิดจากการบังคับ

6. การแนะแนวถือว่าเป็นส่วนหนึ่งของกระบวนการของการศึกษา ดังนั้นการแนะแนวควรสอดแทรกอยู่ในกระบวนการเรียนการสอนของโรงเรียน เพื่อช่วยให้ผู้เรียนแต่ละคนได้มีการพัฒนาตนเองทุกด้านอย่างมีบูรณาการ (integration)

7. การแนะแนวที่มีประสิทธิภาพ ผู้ที่ทำหน้าที่เป็นผู้แนะแนว (counselor) จะต้องเป็นผู้ที่ ได้รับการศึกษาอบรมทางการแนะแนวมาโดยเฉพาะมีทั้งความรู้ (Knowledge) และทักษะ (Skills) ที่เหมาะสมและมีการจัดดำเนินการแนะแนวอย่างมีระบบ (systematical guidance)

8. ผู้ทำงานด้านการแนะแนวจะต้องเป็นผู้ที่มีมนุษยสัมพันธ์ที่ดี มีความเป็นประชาธิปไตย เป็นผู้ที่ยอมรับฟังความคิดเห็นของผู้อื่น และจะต้องเป็นผู้ที่สามารถทำงานร่วมกับผู้อื่นได้เป็นอย่างดี

9. การจัดบริการแนะแนวจะได้ผลดีมีประสิทธิภาพ จะต้องเกิดจากความร่วมมือและความสมัครใจจากบุคลากรทุกฝ่ายในโรงเรียน และผู้เรียนผู้มารับบริการจะต้องมาด้วยความเต็มใจให้ความร่วมมือด้วย

10. ผู้ทำงานด้านการแนะแนวจะต้องเป็นผู้ที่สามารถเก็บรักษาความลับได้ เพราะถ้าเป็นผู้ที่ไม่สามารถเก็บรักษาความลับได้ ก็จะทำให้ผู้เรียนเกิดความรู้สึกไม่ปลอดภัย ทำให้ขาดความไว้วางใจ และไม่ยินดีที่จะมารับความช่วยเหลือ

ด้าน โกศล มีคุณ กล่าวไว้ว่า การแนะแนวเป็นวิชาชีพชั้นสูง เป็นวิชาชีพที่มีองค์ความรู้เป็นของตนเอง มีวิธีการแสวงหาความรู้ที่เชื่อถือได้ มีวิธีดำเนินการหรือเทคนิคเป็นของตนเอง ผู้ประกอบวิชาชีพการแนะแนวจะต้องผ่านการศึกษาอบรม ผ่านกระบวนการฝึกอย่างเข้มข้น เนื่องจากความเป็นวิชาชีพชั้นสูง และเป็นวิชาชีพที่เกี่ยวข้องกับคุณภาพชีวิตของมนุษย์ หรือความเป็นบุคคลที่มีคุณค่าของสังคม การแนะแนวจึงต้องมีปรัชญาการแนะแนวในการดำเนินการ และจะต้องดำเนินการอยู่บนพื้นฐานความรู้และความเชื่อที่สำคัญ ดังต่อไปนี้

1. คนทุกคนมีคุณค่า บุคคลย่อมมีคุณค่าต่อตนเองต่อครอบครัวต่อสังคมประเทศชาติ และต่อสังคมโลก ในฐานะที่จะสร้างคุณประโยชน์ให้แก่ฝ่ายต่างๆ ได้

2. คนทุกคนมีศักดิ์ศรีแห่งความเป็นคน ในฐานะสมาชิกผู้หนึ่งควรได้รับการยอมรับที่จะมีชีวิตอยู่ในสังคม ในโลกอย่างทัดเทียมกับผู้อื่น

3. คนมีความแตกต่างกันควรได้รับการยอมรับในความแตกต่างนั้นทั้งส่วนดีและส่วนร้าย

4. คนควรได้รับการส่งเสริมพัฒนาในทุกด้านตามศักยภาพอย่างเหมาะสม

5. คนควรมีสิทธิตัดสินใจเลือกการพัฒนา แก้ปัญหาของตนเอง ดังนั้นการแนะแนวจึงมีความสำคัญต่อบุคคลหลายฝ่าย

ปรัชญา และแนวคิดข้างต้น สอดคล้องกับนงลักษณ์ ประเสริฐและจรินทร วินทะวิไชย์ ที่กล่าวถึงปรัชญาการแนะแนว (Philosophy of Guidance) โดยเน้นความเชื่ออย่างมีเหตุผล สำหรับใช้เป็นแนวทางการปฏิบัติ ให้ ความช่วยเหลือบุคคล ได้แก่

1. คนเป็นทรัพยากรที่มีคุณค่าสูงกว่าสิ่งใดๆ จึงควรพัฒนาคนให้เจริญเต็มที่ในทุกด้าน

2. คนทุกคนมีความสำคัญและเป็นหน่วยหนึ่งของสังคมทัดเทียมกัน การให้บริการแนะแนว จึงให้โอกาสอย่างเสมอภาคไม่เลือกชั้นวรรณะ

3. คนมีความแตกต่างกัน (Individual Differences) ทุกด้าน ทั้งทางกาย (เช่นอ้วน ผอม แข็งแรง อ่อนแอ สวย ฯลฯ) ทางอารมณ์ (เช่น ใจดี เยือกเย็น สุขุม อารมณ์ร้อน โกรธง่าย ฯลฯ) ทางสังคม (เช่น ชอบอยู่กับคนกลุ่มมาก ชอบอยู่คนเดียว ติดต่อประสานงานได้ดี ฯลฯ) และทางสติปัญญา (เช่น ฉลาด เฉียบ คม คิดเร็ว จำแม่น มีเหตุผล คิดช้า ขาดเหตุผล ฯลฯ)

4. คนทุกคนย่อมมีศักยภาพ (Potentiality) มีความสามารถพิเศษเฉพาะตน รู้จักดูแลและปรับตัวเองให้มีชีวิตรอด ถ้าได้รับการแนะแนวทางที่ถูกต้อง และมีโอกาสเหมาะสมก็จะสามารถช่วยตนเองให้พัฒนาเจริญงอกงามได้ เช่น คนไทยบางคนที่พูดติดอ่าง เมื่ออยู่กับทหารอเมริกันระยะหนึ่ง ก็สามารถสื่อสารภาษาอังกฤษได้คล่องแคล่ว นักร้องลูกทุ่งยอดนิยมหลายคนที่มีการศึกษาน้อย เมื่อได้รับการฝึกวิธีร้องเพลงก็สามารถร้องเพลงได้ชัดเจน ไพเราะ

5. คนทุกคนย่อมมีศักดิ์ศรี (Prestige) มีความดีงาม รักตัวเอง ยอมรับตนเองและต้องการให้ผู้อื่นยอมรับหรือให้เกียรติ ถ้าได้รับการยอมรับก็จะเป็นคนเชื่อมั่นในตนเองและปรับตัวได้

6. คนย่อมมีการเปลี่ยนแปลง (Change) ในช่วงชีวิตคนคนหนึ่ง จะเห็นได้ชัดว่ามีการเปลี่ยนแปลงตั้งแต่เกิดจนถึงตาย โดยเริ่มเปลี่ยนวัย จากวัยทารก เป็นวัยเด็กตอนต้น วัยเด็กตอนกลาง วัยเด็กตอนปลาย วัยรุ่น วัยกลางคนหรือวัยทำงาน วัยผู้ใหญ่ สุดท้ายคือวัยชราซึ่งแต่ละวัยย่อมมีการเปลี่ยนแปลงทางกาย ทางอารมณ์ ทางสังคม และสติปัญญา

7. แต่ละบุคคลย่อมมีปัญหา (Problem) และต้องการจะแก้ปัญหาเศรษฐี ยาจก คนฉลาด คน เก่ง คนแข็งแรง ฯลฯ ต่างก็มีปัญหา และมีวิธีแก้ปัญหาของตน ซึ่งบางคนจำเป็นต้องรับความช่วยเหลือจากผู้อื่น เมื่ออยู่ในสถานการณ์วิกฤต (Critical period) แต่บางคนก็ต้องการความช่วยเหลืออยู่เสมอ

8. พฤติกรรมทุกอย่างย่อมมีสาเหตุ (All Behaviors are Caused) พฤติกรรมเดียวอาจเกิดจาก สาเหตุเดียว หรือหลายสาเหตุ เช่น พฤติกรรมเรียนดี อาจมาจากการตั้งใจเรียนเพียงอย่างเดียว หรือ อาจมาจากการตั้งใจเรียนสม่ำเสมอ การทบทวนบทเรียน การตั้งคำถามตนเอง การค้นคว้าเพิ่มเติม การทำแบบฝึกหัด มีความถนัดพิเศษ ฯลฯ หลายๆ สาเหตุ เป็นต้น

9. บุคคลย่อมต้องการพัฒนาตนเอง บุคคลจะมีความสุข และภูมิใจ เมื่อได้ใช้ความสามารถของตนอย่างเต็มที่ จนสามารถแก้ปัญหาต่างๆ ได้สำเร็จ

2.1.3 ความสำคัญของการแนะแนว

กระทรวงศึกษาการ กล่าวว่า การแนะแนวมีความสำคัญดังนี้

1. เพื่อให้ผู้เรียนค้นพบความถนัด ความสามารถ ความสนใจของตนเอง รักและเห็นคุณค่าในตนเองและผู้อื่น

2. เพื่อให้ผู้เรียนรู้จักการแสวงหาความรู้จากข้อมูล ข่าวสาร แหล่งเรียนรู้ ทั้งด้านการศึกษา อาชีพ ส่วนตัว สังคม เพื่อนำไปใช้ในการวางแผนเลือกแนวทางการศึกษา อาชีพได้เหมาะสมสอดคล้องกับศักยภาพของตนเอง

3. เพื่อให้ผู้เรียนได้พัฒนาบุคลิกภาพและปรับตัวอยู่ในสังคมได้อย่างมีความสุข

4. เพื่อให้ผู้เรียน มีความรู้ มีทักษะ มีความคิดสร้างสรรค์ในงานอาชีพและมีเจตคติที่ดีต่ออาชีพสุจริต

5. เพื่อให้ผู้เรียนมีค่านิยมที่ดีงามในการดำเนินชีวิต เสริมสร้างวินัย คุณธรรมและจริยธรรมแก่ผู้เรียน

6. เพื่อให้ผู้เรียนมีจิตสำนึกในการรับผิดชอบต่อตนเอง ครอบครัว สังคม และประเทศ

2.1.4 ขอบข่ายงานแนะแนว

กระทรวงศึกษาธิการ ได้กล่าวว่า การจัดการแนะแนวมีขอบข่าย 3 ด้าน ดังนี้

1. ด้านการศึกษาให้ผู้เรียนได้พัฒนาตนเองในด้านการเรียนอย่างเต็มศักยภาพ รู้จัก แสวงหา และใช้ข้อมูลประกอบการวางแผนการเรียน หรือการศึกษาต่อได้อย่างมีประสิทธิภาพ มี นิสัยใฝ่รู้ใฝ่เรียน มีวิธีการเรียนรู้ และสามารถวางแผนการเรียน หรือการศึกษาต่อได้อย่างเหมาะสม

2. ด้านอาชีพ ให้นักเรียนได้รู้จักตนเองในทุกด้าน รู้และเข้าใจโลกของงานอาชีพอย่างหลากหลาย มีเจตคติที่ดีต่ออาชีพที่สุจริต มีการเตรียมตัวสู่อาเซียน สามารวางแผนเพื่อประกอบอาชีพที่ตนเองมีความถนัดและความสนใจ

3. ด้านส่วนตัวและสังคม ให้นักเรียนรู้จักและเข้าใจตนเอง รู้และเห็นคุณค่าของตนเองและผู้อื่น มีวุฒิภาวะทางอารมณ์ มีเจตคติที่ดีต่อการมีชีวิตที่ดี มีคุณภาพ มีทักษะชีวิต สามารถปรับตัวและดำรงชีวิตอยู่ในสังคมได้อย่างมีความสุข

2.1.5 หลักการจัดการแนะแนว

กรมวิชาการ ได้กล่าวถึง หลักการจัดการแนะแนวไว้ดังนี้

1. จัดให้กับทุกคน (ไม่เลือกปฏิบัติ) และให้บริการด้วยความเคารพในเกียรติแห่งความเป็นมนุษย์ที่เท่าเทียมกัน โดยคำนึงถึงประโยชน์ของผู้รับบริการเป็นสำคัญ

2. การจัดบริการต้องคำนึงถึงสิทธิเสรีภาพของบุคคลไม่มีการบังคับ

3. การให้บริการต้องเปิดโอกาสให้ทุกคนได้มีบทบาทสำคัญในการใช้ปัญญาเรียนรู้หาวิธีแก้ปัญหาด้วยตนเองและได้พัฒนาตนเต็มตามศักยภาพ

4. การให้บริการต้องตอบสนองตามความต้องการของผู้รับบริการ และคำนึงถึงความแตกต่างระหว่างบุคคล

5. การให้บริการคือการอำนวยเหตุปัจจัยที่เหมาะสมในการสร้างเสริมพัฒนาการหรือพฤติกรรมที่พึงประสงค์

6. การให้บริการช่วยเหลือเพื่อนมนุษย์ เป็นภารกิจที่พึงกระทำด้วยความเมตตา และด้วยความเข้าใจและทุกคนควรมีส่วนร่วม

นอกจากนี้กระทรวงศึกษาธิการยังได้กล่าวถึง หลักการในการจัดการแนะแนวว่า ควรมีการจัดให้สอดคล้องกับสภาพปัญหา ความต้องการ ความสนใจ ธรรมชาติของผู้เรียน และวิสัยทัศน์ของสถานศึกษาที่ตอบสนองจุดมุ่งหมายของหลักสูตรแกนกลางการศึกษาขั้นพื้นฐานพุทธศักราช 2551 ให้ครอบคลุมทั้งด้านการศึกษา อาชีพ ส่วนตัวและสังคม เน้นผู้เรียนเป็นสำคัญ ให้ผู้เรียนมีส่วนร่วมในการปฏิบัติกิจกรรมจนเกิดการเรียนรู้ เกิดทักษะชีวิต โดยมีครูผู้รับผิดชอบกิจกรรมและประสาน ความร่วมมือกับครูหรือผู้มีส่วนเกี่ยวข้อง

2.1.6 แนวทางในการจัดการแนะแนว

กระทรวงศึกษาธิการกล่าวว่า การจัดกิจกรรมแนะแนวต้องสอดคล้องทั้งหลักการและวัตถุประสงค์ มีแนวทางการดำเนินการไว้ดังนี้

1. ศึกษาวิเคราะห์สภาพปัญหา ความต้องการ ความสนใจ และธรรมชาติของผู้เรียน

2. วิเคราะห์สมรรถนะสำคัญของผู้เรียน คุณลักษณะที่พึงประสงค์ วิสัยทัศน์ของสถานศึกษาและข้อมูลของผู้เรียนรายบุคคล

3. กำหนดสัดส่วนกิจกรรมแนะแนวให้ครอบคลุมด้านการศึกษา ด้านอาชีพ ด้านส่วนตัว และสังคม

4. กำหนดวัตถุประสงค์การจัดกิจกรรมแนะแนวของสถานศึกษา

5. ออกแบบการจัดกิจกรรมแนะแนว

6. จัดทำแผนการจัดกิจกรรมแนะแนว

7. จัดกิจกรรมแนะแนวตามแผนและประเมินผลการจัดกิจกรรม

8. ประเมินเพื่อตัดสินผลและสรุปรายงาน

สำหรับการจัดให้มีการประเมินเพื่อตัดสินผล ควรดูจากเวลาเข้าร่วมกิจกรรม การปฏิบัติกิจกรรม ผลงาน/ชิ้นงาน/คุณลักษณะของผู้เรียน ผลการประเมินเป็น “ผ่าน” และ “ไม่ผ่าน หากผ่านตาม เกณฑ์ให้ส่งต่อผลการประเมิน แต่ถ้าไม่ผ่านตามเกณฑ์ให้จัดการซ่อมเสริม

2.1.7 มาตรฐานการแนะแนว

มาตรฐานการแนะแนวที่สอดคล้องกับแผนยุทธศาสตร์การแนะแนว ระดับการศึกษาขั้นพื้นฐานในช่วงแผนพัฒนาเศรษฐกิจและสังคมแห่งชาติฉบับที่ 11 (พ.ศ. 2555 - 2559) เป็นแนวทางให้สถานศึกษา และหน่วยงานที่เกี่ยวข้องได้พัฒนาคุณภาพการแนะแนว และการประเมินงานแนะแนว ประกอบด้วย 3 ด้าน มี 6 มาตรฐาน 14 ตัวบ่งชี้ ดังต่อไปนี้

ด้านที่ 1 คุณภาพผู้เรียน 2 มาตรฐาน 5 ตัวบ่งชี้

- มาตรฐานที่ 1 ผู้เรียนรู้จัก เข้าใจ รักและเห็นคุณค่าในตนเองและผู้อื่น

ตัวบ่งชี้ที่ 1.1 ผู้เรียนรู้จักเข้าใจ รัก และเห็นคุณค่าในตนเอง และพัฒนาตนเองอย่างต่อเนื่อง

ตัวบ่งชี้ที่ 1.2 ผู้เรียนรู้จัก เข้าใจ รักและเห็นคุณค่าผู้อื่น และปฏิบัติตนต่อผู้อื่นอย่างเหมาะสม

- มาตรฐานที่ 2 ผู้เรียนสามารถวางแผนชีวิตด้านการศึกษา ด้านอาชีพ และด้านส่วนตัวและสังคม

ตัวบ่งชี้ที่ 2.1 ผู้เรียนสามารถศึกษาวิเคราะห์ ตัดสินใจ แก้ปัญหาและวางแผนด้านการศึกษา

ตัวบ่งชี้ที่ 2.2 ผู้เรียนสามารถคิดวิเคราะห์ ตัดสินใจ แก้ปัญหาและวางแผนด้านอาชีพ

ตัวบ่งชี้ที่ 2.3 ผู้เรียนสามารถคิดวิเคราะห์ ตัดสินใจ แก้ปัญหาและวางแผนด้านส่วนตัวและสังคม

ด้านที่ 2 คุณภาพการดำเนินงานแนะแนว 3 มาตรฐาน 5 ตัวบ่งชี้

- มาตรฐานที่ 3 ครูแนะแนว/ครูที่ทำหน้าที่แนะแนว ดำเนินงานแนะแนวตามหลักการปรัชญา ขอบข่าย เป้าหมาย และจรรยาบรรณทางการแนะแนว

ตัวบ่งชี้ที่ 3.1 มีความรู้ความสามารถในการดำเนินงานแนะแนวตามปรัชญา ขอบข่าย เป้าหมายการแนะแนว

ตัวบ่งชี้ที่ 3.2 ปฏิบัติตามจรรยาบรรณการแนะแนว

- มาตรฐานที่ 4 สถานศึกษาจัดบริการแนะแนวและส่งเสริมระบบดูแลช่วยเหลือนักเรียน

ตัวบ่งชี้ที่ 4.1 จัดบริการแนะแนวทั้ง 5 บริการ ครอบคลุมขอบข่ายการแนะแนวอย่างเป็นระบบและต่อเนื่อง

ตัวบ่งชี้ที่ 4.2 ส่งเสริมระบบดูแลช่วยเหลือนักเรียน

- มาตรฐานที่ 5 สถานศึกษาจัดกิจกรรมแนะแนวตามหลักสูตรสถานศึกษา

ตัวบ่งชี้ที่ 5.1 จัดกิจกรรมแนะแนวสอดคล้องกับบริบทของสถานศึกษาและครอบคลุมขอบข่ายการแนะแนว

ด้านที่ 3 คุณภาพการบริหารจัดการแนะแนว 1 มาตรฐาน 4 ตัวบ่งชี้

- มาตรฐานที่ 6 สถานศึกษามีการบริหารจัดการแนะแนวอย่างมีคุณภาพ

ตัวบ่งชี้ที่ 6.1 ผู้บริหารให้การสนับสนุนการดำเนินงานแนะแนว

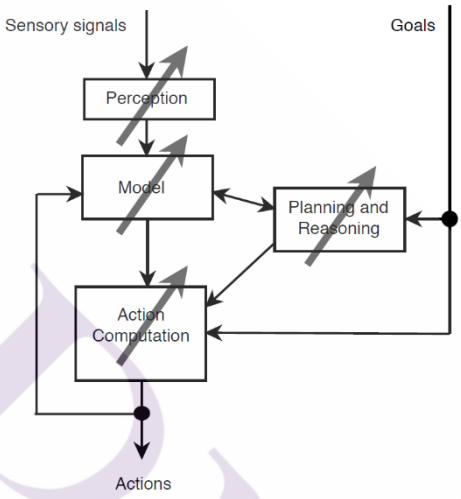
ตัวบ่งชี้ที่ 6.2 บริหารจัดการงานแนะแนวอย่างเป็นระบบ

ตัวบ่งชี้ที่ 6.3 จัดให้มีครูแนะแนว/ครูที่ทำหน้าที่แนะแนว

ตัวบ่งชี้ที่ 6.4 มีภาคีเครือข่ายเข้ามามีส่วนร่วมในงานแนะแนว

2.2 การเรียนรู้ของเครื่อง (Machine Learning)

Machine Learning ศาสตร์ของการให้คอมพิวเตอร์ทำงานหรือเกิดกระบวนการเรียนรู้ได้ด้วยตนเองของเครื่องจากตัวอย่าง โดยไม่ได้มีการป้อนคำสั่งแต่ได้ผลลัพธ์ที่มีค่าความถูกต้องประสิทธิภาพสูง การเรียนรู้ของเครื่องมักจะหมายถึงการเปลี่ยนแปลงในระบบทำงานที่เกี่ยวข้องกับปัญญาประดิษฐ์ (AI) งานดังกล่าวเกี่ยวข้องกับการรับรู้ การวินิจฉัย การวางแผน การควบคุมหุ่นยนต์ การทำนาย อาจเป็นการเพิ่มประสิทธิภาพให้กับระบบที่ดำเนินการอยู่ หรือการสังเคราะห์ระบบใหม่

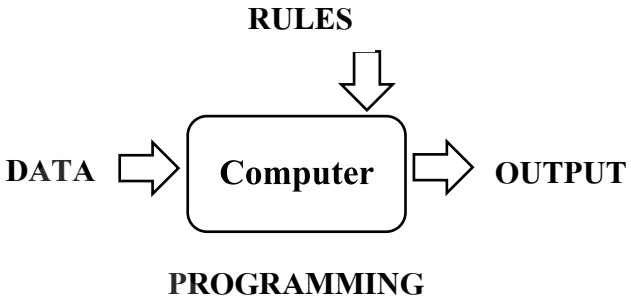


ภาพที่ 2.1 ระบบ AI

ภาพที่ 2.1 แสดงสถาปัตยกรรมของตัวแทน AI เป็นการรับรู้การจำลองสภาพแวดล้อมและคำนวณการกระทำที่เหมาะสม โดยคาดการณ์ถึงผลกระทบการเปลี่ยนแปลงที่เกิดกับส่วนประกอบในระบบ AI อาจนับเป็นการเรียนรู้ กลไกการเรียนรู้ที่แตกต่างกันจะถูกนำมาใช้ขึ้นอยู่กับระบบย่อยที่มีการเปลี่ยนแปลง เราจะศึกษาวิธีการเรียนรู้ต่างๆ มากมายใน Machine Learning

2.2.1 ความแตกต่างระหว่าง Machine Learning และการเขียนโปรแกรมแบบเดิม

การเขียนโปรแกรมแบบเดิมแตกต่างกับ Machine Learning การเขียนโปรแกรมแบบเดิมนั้นการทำงานทั้งหมดจะต้องถูกกำหนดแนวทางไว้ชัดเจนด้วยกฎจากโปรแกรมเมอร์เองซอฟต์แวร์จะถูกพัฒนาขึ้น โดยแต่ละกฎจะขึ้นอยู่กับพื้นฐานความเข้าใจด้านตรรกศาสตร์ (Logic Foundation) การทำงานและส่งผลลัพธ์ออกมาตามคำสั่งตามตรรกะ (Logical statement) เมื่อระบบเริ่มซับซ้อนมากขึ้นยิ่งจำเป็นต้องมีกฎมากขึ้นการเขียนคำสั่งก็มากขึ้นดังภาพที่ 2.2



ภาพที่ 2.2 คำสั่งตามตรรกะแบบเดิม (Logical statement)

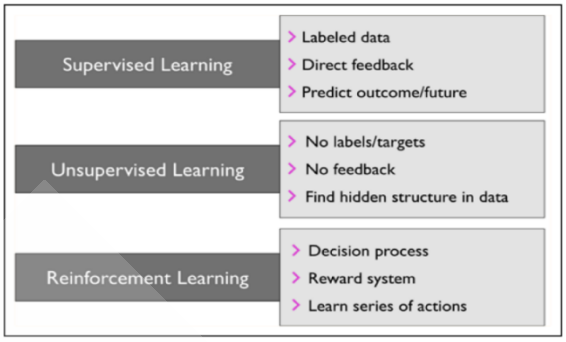
Machine Learning มีวิธีการเขียนโปรแกรมสะดวกกว่าแบบสมัยก่อนโดยเครื่อง (Machine) จะเรียนรู้ว่าข้อมูลขาเข้าและข้อมูลขาออกเกี่ยวข้องกับอย่างไรและรู้ว่าจะเขียนกฎหนึ่งขึ้นมาอย่างไร โปรแกรมเมอร์ไม่จำเป็นต้องเขียนกฎใหม่ทุกครั้งที่มีข้อมูลใหม่ Algorithms จะปรับวิธีเข้ากับข้อมูลใหม่เพื่อปรับปรุงประสิทธิภาพโดยอัตโนมัติ

2.2.2 การทำงานของ Machine Learning

วิธีที่ Machine เรียนรู้คล้ายกับมนุษย์ คือเรียนรู้จากประสบการณ์ ยิ่งมีชุดเรียนรู้มากยิ่งพยากรณ์ได้ดีขึ้น โดยเมื่อมีประสบกับเหตุการณ์ที่ไม่เคยเจอมาก่อนก็จะมีความเป็นไปได้ที่จะเรียนรู้และคำนวนได้ดี Machine สามารถถูกฝึกได้ เพื่อที่จะเพิ่มความแม่นยำในการพยากรณ์ Machine สามารถมองเห็นแบบที่เราต้องการให้เห็นได้ เมื่อป้อนตัวอย่างที่เหมือนกันให้ Machine ก็จะสามารถค้นหาผลลัพธ์จนพบ ข้อสังเกตหากถูกสั่งให้ค้นหาในสิ่งที่ไม่เคยถูก Training มาก่อนก็จะไม่ประสบความสำเร็จ

2.2.3 ประเภทการเรียนรู้ของ Machine Learning

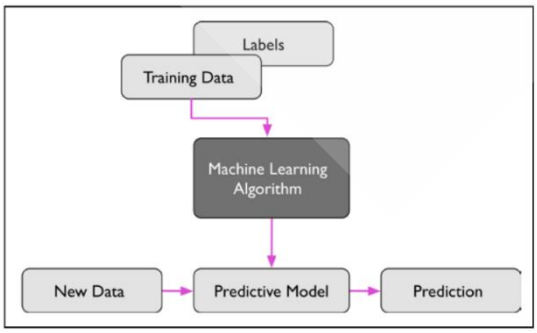
การเรียนรู้ของ Machine Learning นิยมแบ่งออกเป็นการเรียนรู้ได้ 3 แบบ ได้แก่ 1. การเรียนรู้แบบมีผู้สอน(Supervised Learning) 2. การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) และ3. การเรียนรู้แบบเสริมแรง (Reinforcement Learning) ความแตกต่างพื้นฐานระหว่างการเรียนรู้ทั้งสามประเภทและการใช้ตัวอย่างแนวคิดการแก้ปัญหาในทางปฏิบัติ ดังภาพที่ 2.3



ภาพที่ 2.3 ประเภทการเรียนรู้ของ Machine Learning

1. การเรียนรู้แบบมีผู้ดูแล (Supervised learning)

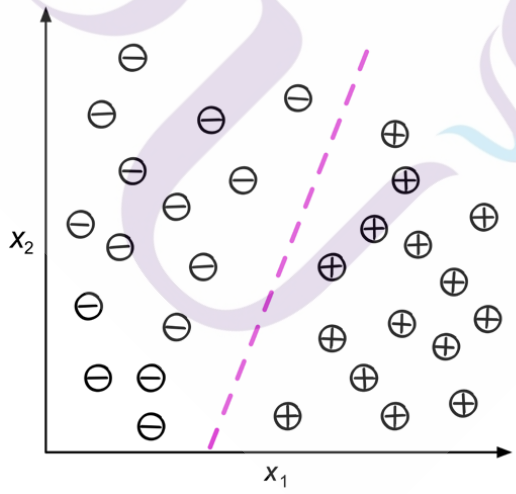
เป้าหมายหลักใน Supervised learning คือ การเรียนรู้รูปแบบจากข้อมูลการฝึกอบรมที่มี Labels ช่วยให้เราสามารถคาดเดาเกี่ยวกับข้อมูลที่มองไม่เห็นหรือข้อมูลที่จะเข้ามาในอนาคต ภายใต้การดูแลหมายถึงชุดของตัวอย่างที่มี Labels ที่ต้องการและAlgorithms จำเป็นต้องใช้ข้อมูลสำหรับ Training data และส่วนที่รับกลับมาเพื่อปรับปรุง (Feedback) เพื่อที่จะเรียนรู้ความสัมพันธ์ระหว่างข้อมูลที่ถูกป้อนเข้ามาสู่ข้อมูลที่ออกไป เมื่อพิจารณาจากตัวอย่างของการกรองอีเมลขยะ เราสามารถฝึกโมเดลโดยใช้ไฟล์ Algorithm Machine Learning ภายใต้การดูแลของอีเมลที่มี Labels อีเมลนั้น ถูกทำเครื่องหมายอย่างถูกต้องว่าเป็นสแปมหรือไม่ใช่สแปมเพื่อคาดเดาว่าเป็นอีเมลใหม่หรือไม่เป็นอย่างใดอย่างหนึ่งจากสองประเภท งานการเรียนรู้ภายใต้การดูแลพร้อม Labels ชั้นเรียนแยกกัน เช่นในตัวอย่างการกรองอีเมลขยะก่อนหน้านี้เรียกอีกอย่างว่า การจัดประเภทงาน อีกประเภทย่อยของการเรียนรู้ภายใต้การดูแลคือ การถดถอยซึ่งผลลัพธ์สัญญาณเป็นค่าต่อเนื่อง ดังภาพที่ 2.4



ภาพที่ 2.4 การเรียนรู้แบบมีผู้ดูแล (Supervised learning)

1.1 การแบ่งแยกประเภท (Classification)

การแบ่งแยกประเภทจัดเป็นหมวดหมู่หนึ่งของการเรียนรู้แบบมีผู้ดูแล (Supervised learning) โดยมีเป้าหมายเพื่อคาดการณ์ Labels กำกับระดับหมวดหมู่ของ Instances โดยอิงจากการสังเกตในอดีต Labels คลาสเหล่านั้นเป็นค่าที่ไม่ต่อเนื่อง และไม่เรียงลำดับซึ่งสามารถเข้าใจได้ในฐานะกลุ่มสมาชิกของ Instances เช่น การตรวจหาสแปมอีเมล การจัดประเภทไบนารี โดยที่ Algorithms ของ Machine Learning จะเรียนรู้ชุดของกฎเพื่อแยกแยะระหว่างสองคลาสที่เป็นไปได้ ระหว่างอีเมลขยะและอีเมลที่ไม่ใช่สแปม อย่างไรก็ตาม ชุดของ Labels คลาสไม่จำเป็นต้องมีลักษณะไบนารี โมเดลการคาดการณ์ที่เรียนรู้โดย Algorithms การเรียนรู้ภายใต้การดูแลสามารถกำหนด Labels กำกับคลาสที่แสดงในชุดข้อมูลการ Training ให้กับ Instances ที่ไม่มี Labels กำกับ ตัวอย่างทั่วไปของงานการจำแนกประเภทหลายคลาส คือการรู้จำอักขระที่เขียนด้วยลายมือ ที่นี่ เราสามารถรวบรวมชุดข้อมูลการ Training ที่ประกอบด้วยตัวอย่างที่เขียนด้วยลายมือหลายตัวของแต่ละตัวอักษรในตัวอักษร หากผู้ใช้ระบุอักขระที่เขียนด้วยลายมือใหม่ผ่านอุปกรณ์ป้อนข้อมูล โมเดลการทำนายของเราจะทำนายตัวอักษรที่ถูกต้องในตัวอักษรได้อย่างแม่นยำ อย่างไรก็ตาม Machine Learning ของเราไม่สามารถ จดจำตัวเลขตั้งแต่ 0 ถึง 9 ได้อย่างถูกต้องหากไม่ได้เป็นส่วนหนึ่งของชุดข้อมูลที่ถูกTraining

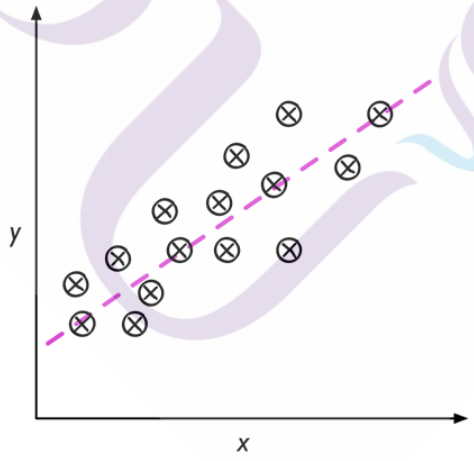


ภาพที่ 2.5 การจำแนก Classification

จากภาพที่ 2.5 เป็นแนวคิดของการจำแนกเลขฐานสองที่ได้รับตัวอย่างการ Training 30 ตัวอย่างโดยตัวอย่างการ Training 15 ตัวอย่าง ถูกระบุว่าเป็นคลาสเชิงลบ (เครื่องหมายลบ) และตัวอย่างการ Training 15 ตัวอย่าง ถูกระบุว่าเป็นคลาสบวก (เครื่องหมายบวก) ในสถานการณ์สมมตินี้ ชุดข้อมูลของเราเป็นแบบสองมิติ ซึ่งหมายความว่าแต่ละตัวอย่างมีค่าสองค่าที่เกี่ยวข้อง $\square_1$ และ $\square_2$ ตอนนี้ เราสามารถใช้Algorithm ใน Machine Learning แบบ Supervised learning เพื่อเรียนรู้กฎขอบเขตการตัดสินใจที่แสดงเป็นเส้นประที่สามารถแยกสองคลาสเหล่านั้นและจัดประเภทข้อมูลใหม่เป็นแต่ละหมวดหมู่ของสองหมวดหมู่โดยมีค่า $\square_1$ และ $\square_2$

1.2 การถดถอย (Regression)

การวิเคราะห์การถดถอยอาจเรียกว่าการทำนายผลลัพธ์อย่างต่อเนื่อง การวิเคราะห์การถดถอยเกิดขึ้นเมื่อเราได้รับตัวแปรทำนายจำนวนหนึ่ง (Explanatory) และตัวแปรตอบสนองอย่างต่อเนื่อง (Outcome or Target) และเราพยายามค้นหาความสัมพันธ์ระหว่างตัวแปรเหล่านั้นที่ทำให้เราสามารถทำนายผลลัพธ์ได้เช่น การทำนายคะแนนคณิตศาสตร์ของนักเรียน หากมีความสัมพันธ์ระหว่างเวลาที่ใช้ในการศึกษาสำหรับการทดสอบกับคะแนนสุดท้าย เราอาจใช้เป็นข้อมูลการ Training เพื่อเรียนรู้แบบจำลองที่ใช้เวลาศึกษาเพื่อทำนายคะแนนการทดสอบของนักเรียนในอนาคตที่กำลังวางแผนในการทำแบบทดสอบนี้



ภาพที่ 2.6 การถดถอยเชิงเส้น

จากภาพที่ 2.6 แสดงแนวคิดของการถดถอยเชิงเส้นจากตัวแปรทำนาย x และตัวแปรตอบสนอง y เมื่อจัดเส้นตรงกับข้อมูลนี้ที่ลดระยะทาง ซึ่งโดยปกติเป็นระยะห่างกำลังสองเฉลี่ยระหว่างกลุ่มตัวอย่าง และเส้นตอนนี้เราสามารถใช้การสกัดกั้นและความชันที่เรียนรู้จากข้อมูลนี้เพื่อคาดการณ์ตัวแปรผลลัพธ์ของข้อมูลใหม่ได้

2. การเรียนรู้แบบไม่มีผู้สอน (Unsupervised learning)

ในการเรียนรู้แบบไม่มีผู้ดูแลเรากำลังจัดการกับข้อมูลที่ไม่มี Labels หรือข้อมูลโครงสร้างที่ไม่รู้จัก การใช้การเรียนรู้แบบ Unsupervised learning เป็นเทคนิคที่สามารถสำรวจโครงสร้างของข้อมูล เพื่อดึงข้อมูลที่มีความหมายโดยไม่ต้องแนะนำตัวแปรผลลัพธ์ที่รู้จักหรือ Reward function

2.2.4 ขั้นตอนการทำ Machine Learning ของ Aurélien Géron มี 8 ขั้นตอนดังนี้

1) การมองภาพรวมและการวางกรอบปัญหา (Look at the big picture)

เริ่มต้นจากการกำหนดวัตถุประสงค์ในการที่จะแก้ปัญหา โดยการสำรวจแนวทางการแก้ปัญหาว่ามีวิธีการแก้ปัญหาด้วยวิธีใดบ้างที่จะเหมาะสมกับชุดข้อมูลที่มี เช่น การกำหนดว่าจะใช้วิธีการใด Supervised learning, Unsupervised learning, Reinforcement learning นอกจากนี้ให้เริ่มศึกษาวิธีวัดประสิทธิภาพที่สอดคล้องและเหมาะสมกับวัตถุประสงค์ อาจศึกษาจากกรณีศึกษาที่ใกล้เคียงหรือคล้ายคลึงกับปัญหาหรือความต้องการที่กำหนด ต่อมากำหนดทรัพยากรด้านต่างๆ ที่จำเป็นต้องใช้และอาจทำการทดสอบสมมุติฐานที่มีทั้งหมด

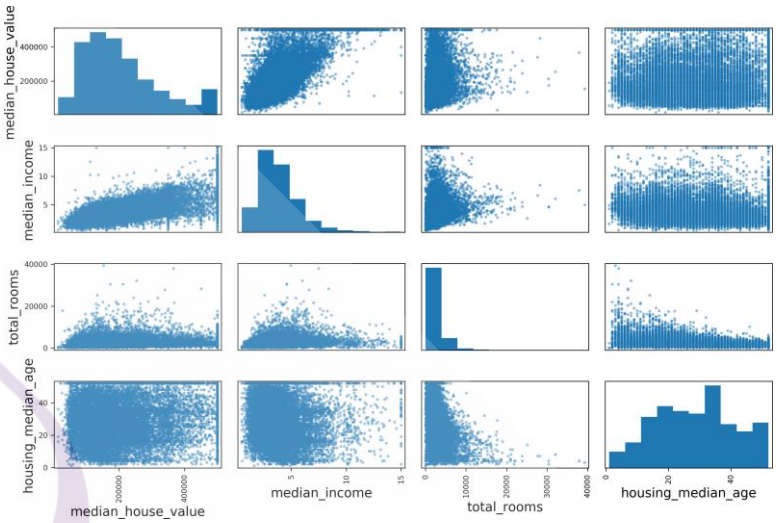
2) การรับและการรวบรวมข้อมูล (Get the data)

การจัดการข้อมูลอย่างเป็นระบบตั้งแต่กระบวนการนำเข้าข้อมูลจนถึงการแปลงและจัดเก็บรวบรวมข้อมูลให้เป็นระบบจะทำให้การทำ Machine Learning ได้สะดวกขึ้น โดยเริ่มจากการจัดทำรายการข้อมูลที่ต้องใช้ แสวงหาแหล่งข้อมูลที่จำเป็น ตรวจสอบความถูกต้องของแหล่งที่มาทั้งทางด้านข้อตกลงทางด้านลิขสิทธิ์และกฎหมายที่เกี่ยวข้องต่างๆ ทำการรวบรวมและแปลงข้อมูลในรูปแบบที่จัดกระทำได้ง่าย อาจจัดกระทำในรูปแบบ Labels เพื่อให้ง่ายต่อการสร้าง Machine Learning รวมทั้งกำหนดข้อมูลชุดฝึกและชุดทดสอบ

3) การสำรวจและการค้นพบข้อมูลเชิงลึก (Discover and visualize the data to gain insights)

การตรวจสอบให้แน่ใจว่าเรากำลังทำข้อมูลชุดการฝึกเท่านั้น ตรวจสอบขนาดของชุดการฝึก หากชุดฝึกมีขนาดใหญ่มาก คุณอาจต้องการสุ่มตัวอย่างเพื่อให้การปรับเปลี่ยนทำได้ง่ายและรวดเร็ว หากชุดฝึกมีขนาดไม่ใหญ่มาก เราก็อาจจะทำการดำเนินงานได้ทันที อาจเริ่มจากการหาค่าสัมประสิทธิ์สหสัมพันธ์มาตรฐานของชุดข้อมูลต่างๆ หรืออาจใช้อีกวิธีในการตรวจสอบความสัมพันธ์ระหว่างAttributes คือการใช้ฟังก์ชัน scatter_matrix ซึ่งจะพล็อต Attributes ที่เป็น

ตัวเลขทั้งหมดเทียบกับAttribute ตัวเลขอื่นๆ โดยเน้นที่Attributes ที่มีแนวโน้มว่าจะสัมพันธ์กับมูลค่า ดังภาพที่ 2.7



ภาพที่ 2.7 Scatter matrix

4) จัดเตรียมข้อมูลเพื่อขั้นตอนวิธีการเรียนรู้ของเครื่อง (Prepare the data for Machine Learning algorithms)

การเตรียมข้อมูลสำหรับ Machine Learning algorithms โดยทำการเขียนฟังก์ชัน การสร้างไลบรารีของฟังก์ชัน การใช้ฟังก์ชันเพื่อแปลงข้อมูลใหม่ก่อนที่จะป้อนไปยัง Algorithms เมื่อเราสามารถลองใช้การแปลงข้อมูลแบบต่าง ๆ ได้เราก็จะสามารถทดลองและทราบได้ว่าวิธีการใดดีที่สุดกับชุดข้อมูลของเรา ในขั้นตอนนี้รวมทั้งการทำ Data Cleaning การทำ Feature Selection และการทำ Feature Scaling โดยการทำ Standardize หรือ Normalize เพื่อให้ข้อมูลแต่ละ Feature อยู่ใน Scale ที่ใกล้เคียงกัน

5) การคัดสรรและการฝึกแบบจำลอง (Select a model and training it)

เริ่มด้วยการพิจารณาขนาดของข้อมูลเพื่อกำหนดขั้นตอนการทำ Training Set โดยการทำ Sample Training Set กับ Model หลายๆแบบ ด้วย Algorithms ที่พิจารณาแล้วว่าเหมาะสมกับลักษณะของชุดข้อมูล เช่น การใช้ Decision Tree ,Naïve Bayes ,Artificial Neural Network ,Support Vector Machine โดยการวิเคราะห์วัดประสิทธิภาพเปรียบเทียบด้านต่างๆ และค่า Error ของแต่ละรูปแบบทำการทดลองซ้ำจนค่าของข้อมูลเกิดการอิ่มตัว และทำการคัดเลือก Model

ที่มีประสิทธิภาพดีในแต่ละด้านเพื่อนำมาเลือกตามความเหมาะสมตามวัตถุประสงค์ของงานอีกครั้งหนึ่ง

6) การปรับปรุงแบบจำลอง (Fine-tune your model)

เมื่อเราได้ Model ที่มีประสิทธิภาพดีแล้วในขั้นตอนนี้เราต้อง ปรับแต่งให้มีค่าประสิทธิภาพดีขึ้น ปรับแก้เพื่อหาค่า Hyper parameter ที่รวมกันเป็นอย่างดีที่มีค่าสูงที่สุด และข้อสังเกตคือ ต้องไม่เป็นการปรับแต่งจนเกิดการกำหนดวิธีที่เป็นการเฉพาะกับข้อมูลบางชุดมากเกินไปจนเป็น Over fitting ของวิธีที่มีต่อชุดข้อมูลนี้

7) การนำเสนอวิธีการแก้ปัญหา (Present your solution)

เมื่อปรับปรุง Model จนมีประสิทธิภาพดีเหมาะสมกับเทคนิคที่ใช้แล้ว จัดทำรายงาน และทำสื่อเพื่อการนำเสนอโดยอาจเริ่มจาก การมองภาพรวมและการวางกรอบปัญหา (Look at the big picture) แล้วจึงนำเสนอวิธีการแก้ปัญหาทางเลือกข้อดี ข้อจำกัดในแต่ละวิธี

8) การเริ่มต้นโครงการ การสังเกตการณ์ และการดูแลปรับปรุงระบบ (Launch, monitor, and maintain your system)

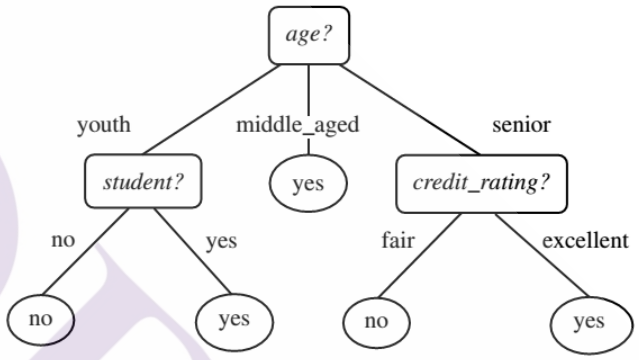
เมื่อ Machine Learning พร้อมสำหรับการใช้งานจริง เริ่มต้นในการใช้งานเราต้องมีกระบวนการสังเกต ตรวจสอบและติดตามประสิทธิภาพของ Machine Learning จากค่าประสิทธิภาพ เมื่อค่าเริ่มลดลงต้องหาสาเหตุ และแก้ไขในจุดบกพร่อง หรือมีการปรับปรุง Machine เพื่อให้มีความเหมาะสมกับชุดข้อมูลอย่างสม่ำเสมอ การตรวจสอบปัจจัยมีความสำคัญอย่างยิ่งกับค่าสิทธิภาพ Machine ควรมีระบบการเรียนรู้ชุดข้อมูลให้มีประสิทธิภาพด้วยวิธีออนไลน์เพื่อคอยตรวจสอบคุณภาพของข้อมูลที่เข้ามาด้วย ซึ่งเป็นเรื่องสำคัญหากระบบเราเป็นแบบ Online learning

2.3 ขั้นตอนวิธี (Algorithms) ในงานวิจัยนี้เลือกใช้ 4 วิธีดังนี้

2.3.1 Decision Trees

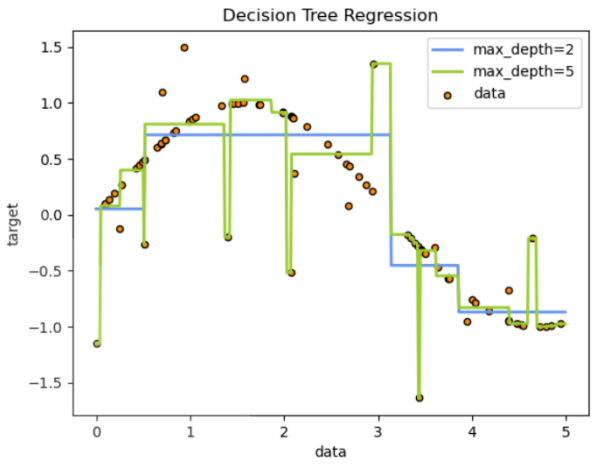
การจำแนกข้อมูลด้วย Decision Trees จะเป็นกระบวนการสร้างต้นไม้ขึ้นเพื่อใช้ในการตัดสินใจจากข้อมูลที่มีหมวดหมู่ข้อมูลแนบอยู่ประกอบไปด้วยโหนดต่างๆ ที่ไม่ใช่โหนดใบ (non-leaf node) ใช้ในการแสดงถึงเงื่อนไขหรือ Attributes ของข้อมูล โดยที่แต่ละกิ่งก้านของแต่ละโหนดนั้นจะหมายถึงค่าที่เป็นไปได้จากการทดสอบกับ Attributes และจะประกอบไปด้วยโหนดใบ (leaf node) ที่จะมีหมวดหมู่ข้อมูลจัดเก็บอยู่ ยกตัวอย่างการทำนายคุณลักษณะของลูกค้าที่จะทำการซื้อคอมพิวเตอร์จากร้านขายอุปกรณ์ไฟฟ้า โดยโหนดต่างๆที่ไม่ใช่โหนดใบจะถูกแทนด้วยสี่เหลี่ยม และโหนดใบจะถูกแทนด้วยวงรี ตามลำดับ จากรูปเราจะเห็นว่าโหนดใบจะเป็นโหนดที่บ่งบอกถึงข้อมูลหมวดหมู่ของคำตอบที่เราต้องการอาทิ เช่น “yes” หมายถึง ลูกค้าจะซื้อคอมพิวเตอร์ และ

“no” หมายถึงลูกค้าจะไม่ซื้อคอมพิวเตอร์ โดย Trees ที่ถูกสร้างขึ้นอาจมีลักษณะเป็นไบนารี่หรืออาจจะไม่เป็นไบนารี่ก็ได้ หลังจากทำการสร้าง Decision Trees แล้ว เราจะสามารถจำแนกข้อมูลได้โดยการจำแนกหมวดหมู่ของข้อมูลเรคคอร์ดใดๆที่ประกอบไปด้วย Attributes ต่างๆ แต่เราจะไม่ทราบหมวดหมู่ข้อมูลในเรคคอร์ดนั้นด้วยการเปรียบเทียบ Attributes ที่อยู่ในโหนดรากกับค่าของ Attributes ในเรคคอร์ดที่พิจารณา โดยจะทำการเปรียบเทียบจากโหนดรากไปจนถึงโหนดใบ เมื่อเราทราบถึงโหนดใบจะทำให้เราทราบถึงหมวดหมู่ข้อมูลของเรคคอร์ดที่ทำการพิจารณา ดังภาพที่ 2.8



ภาพที่ 2.8 การจำแนกคุณลักษณะของลูกค้าที่ทำการซื้อคอมพิวเตอร์

Decision Trees (DTs) ในงานวิจัยนี้ใช้ Scikit-learn โดยเป็นวิธีการเรียนรู้ภายใต้การดูแลแบบไม่มีพารามิเตอร์ที่ใช้สำหรับการจำแนกประเภทและการถดถอย เป้าหมายคือ การสร้างแบบจำลองที่คาดการณ์ค่าของตัวแปรเป้าหมายโดยการเรียนรู้กฎการตัดสินใจอย่างง่ายที่อนุมานจากคุณลักษณะข้อมูล สามารถมองต้นไม้เป็นการประมาณค่าคงที่ทีละส่วนได้ ตัวอย่างเช่น ในตัวอย่างด้านล่าง แผนผังการตัดสินใจเรียนรู้จากข้อมูลแบบเส้นโค้งด้วยชุดของกฎการตัดสินใจแบบ if-then-else ยิ่งต้นไม้ยิ่งลึก กฎการตัดสินใจก็ยิ่งซับซ้อนมากขึ้น และโมเดลยิ่งมีความเหมาะสมมากขึ้นเท่านั้นดังภาพที่ 2.9



ภาพที่ 2.9 กฎการตัดสินใจแบบ if-then-else

ที่มา: https://scikit-learn.org/stable/modules/tree.html#classification

ข้อดีของการใช้ Decision Trees (DTs) ด้วย Scikit-learn

1) สามารถมองเห็นความสัมพันธ์ระหว่างตัวแปรทำให้เข้าใจและตีความได้

2) การเตรียมข้อมูลไม่ยุ่งยากซับซ้อน

3) การสร้างตัวแปรแบบจำลองและลบค่าว่างออกทำได้ง่าย

4) ใช้ทรัพยากรในการทำนายน้อย เช่น ข้อมูลการทำนาย

5) สามารถจัดการข้อมูลทั้งตัวเลขและหมวดหมู่ อย่างไรก็ตามการใช้งาน

6) สามารถจัดการกับปัญหาหลายเอาต์พุต

7) สามารถสร้างโมเดลทำนายแบบสองค่า White box และBlack box โดย White boxใช้ในสถานการณ์ที่กำหนดสามารถสังเกตได้ในแบบจำลองเงื่อนไขนั้นสามารถอธิบายได้ง่ายโดยใช้ตรรกะบูลีน หากเกิดในทางตรงกันข้ามก็แทนในรูปแบบ Black box

8) สามารถที่จะตรวจสอบแบบจำลองโดยใช้การทดสอบทางสถิติทำให้สามารถพิจารณาความน่าเชื่อถือของแบบจำลองได้

9) ทำนายได้ดีแม้ว่าสมมติฐานจะถูกละเมิดโดยแบบจำลองจริงที่ใช้สร้างข้อมูล

ข้อจำกัดของ Decision Trees (DTs) ด้วย Scikit-learn

1) หากใส่ตัวแปรจำนวนมากเพื่อนำมาสร้างแผนภูมิที่ซับซ้อนเกินไปจะทำให้ไม่สามารถสรุปข้อมูลได้ดี

2) แผนผังการตัดสินใจอาจไม่เสถียรเนื่องจากการเปลี่ยนแปลงเล็กน้อยในข้อมูลอาจส่งผลให้มีการสร้าง Trees ที่แตกต่างกัน ปัญหานี้แก้ได้โดยใช้แผนผังการตัดสินใจภายใน

3) การคาดคะเนของแผนผังการตัดสินใจนั้นไม่ราบรื่นหรือต่อเนื่อง อาจเป็นการประมาณค่าคงที่ทีละส่วนตามที่เห็นในรูปด้านบน

4) ปัญหาของการเรียนรู้โครงสร้างการตัดสินใจที่เหมาะสม NP-complete โดยAlgorithms การเรียนรู้เชิงโครงสร้างการตัดสินใจที่ใช้งานได้จริงจะขึ้นอยู่กับฮิวริสติก

5) ปัญหาความเท่าเทียมกันหรือปัญหามัลติเพล็กเซอร์ แผนผังการตัดสินใจไม่สามารถอธิบายได้ เช่น ตรรกศาสตร์ XOR

Decision Tree Classifier

เป็นคลาสที่สามารถทำการจำแนกประเภทหลายคลาสบนชุดข้อมูล Decision Tree Classifier ใช้เป็นอินพุตสองอาร์เรย์ โดยอาร์เรย์ X จะกระจายหรือหนาแน่น (n_samples, n_features) เพื่อเก็บตัวอย่างการฝึก และอาร์เรย์ Y ของค่าจำนวนเต็ม (n_samples) ถือ Labels คลาสสำหรับตัวอย่างการฝึกอบรม

```
from sklearn import tree
X = [[0, 0], [1, 1]]
Y = [0, 1]
clf = tree.DecisionTreeClassifier()
clf = clf.fit(X, Y)
```

ภาพที่ 2.10 Decision Tree Classifier

หลังจากติดตั้งแล้วสามารถใช้แบบจำลอง เพื่อทำนายคลาสของตัวอย่างได้ในกรณีที่มีหลายคลาสที่มีความน่าจะเป็นเท่ากัน ตัวแยกประเภทจะทำนายคลาสที่มีดัชนีต่ำสุดในบรรดาคลาสเหล่านั้นทางเลือกอื่นนอกเหนือจากคลาสเฉพาะ ความน่าจะเป็นของแต่ละคลาสสามารถคาดการณ์ได้ ซึ่งเป็นเศษส่วนของตัวอย่างการฝึกของชั้นเรียนในโหนดใบ

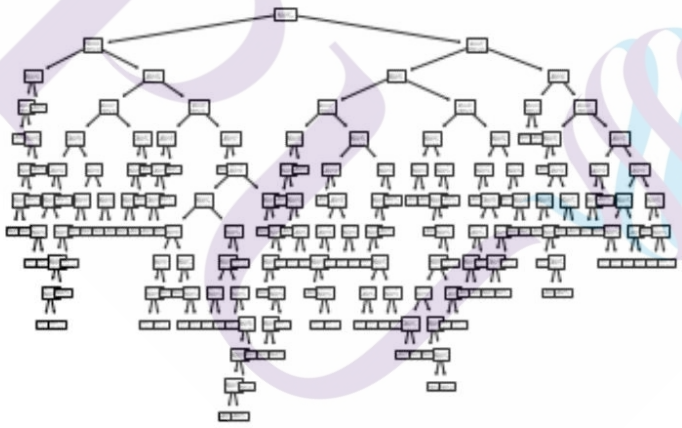
```
    clf.predict_proba([[2., 2.]])
array([[0., 1.]])
```

ภาพที่ 2.11 การทำนายคลาส

Decision Tree Classifier สามารถจัดหมวดหมู่ได้ทั้งแบบไบนารีโดยที่ Labels คือ [-1, 1] และการจัดประเภทแบบหลายคลาสโดยที่Labels คือ [0, …, K-1] การใช้ชุดข้อมูล Iris เราสามารถสร้างต้นไม้ได้ดังภาพที่ 2.13

```
from sklearn.datasets import load_iris
from sklearn import tree
iris = load_iris()
X, y = iris.data, iris.target
clf = tree.DecisionTreeClassifier()
```

ภาพที่ 2.12 การจัดประเภทแบบหลายคลาส



ภาพที่ 2.13 การพล็อต Tree ด้วยฟังก์ชัน plot_tree: tree.plot_tree(clf)

Tree Algorithms

ID3 (Iterative Dichotomiser 3) ได้รับการพัฒนาในปี 1986 โดย Ross Quinlan เป็น Algorithms ที่สร้าง Multiway tree โดยเริ่มจากค้นหาแต่ละโหนด เช่นวิธี Greedy approach ลักษณะหมวดหมู่ที่จะใช้ข้อมูลที่ใหญ่ที่สุดสำหรับเป้าหมายที่เป็นหมวดหมู่ โดยจะสร้างต้นไม้

เติบโตจนใหญ่ที่สุด จากนั้นจึงใช้ขั้นตอนการตัดแต่งกิ่งเพื่อปรับปรุงความสามารถของต้นไม้เพื่อใช้ในการสรุปข้อมูลที่มองไม่เห็น

C4.5 เป็น Algorithms ที่พัฒนาต่อยอดมาจาก ID3 โดยทำการแก้ข้อจำกัดการจัดหมวดหมู่โดยการกำหนด Attributes ที่ไม่ต่อเนื่องแบบไดนามิกตามตัวแปรตัวเลขที่แบ่งพาร์ติชันค่า Attributes แบบต่อเนื่องเป็นชุดของช่วงเวลาที่ไม่ต่อเนื่องกัน C4.5 แปลงแผนผังที่ฝึกแล้ว เช่น เอาต์พุตของ ID3 เป็นชุดของกฎ if-then ความถูกต้องของกฎแต่ละข้อเหล่านี้จะได้รับการประเมินเพื่อกำหนดลำดับที่ควรเป็นเพื่อการตัดแต่งกิ่งทำได้โดยการเอาเงื่อนไขเบื้องต้นของกฎออกหากความถูกต้องของกฎดีขึ้นก็ไม่มีการปรับปรุง

C5.0 เป็นเวอร์ชันล่าสุดของ Quinlan ภายใต้ลิขสิทธิ์เฉพาะใช้หน่วยความจำน้อยกว่าและสร้างชุดกฎที่เล็กกว่า C4.5 ในขณะที่มีความแม่นยำมากขึ้น

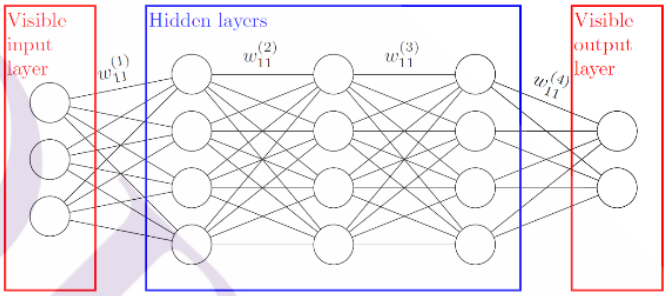
CART (Classification and Regression Trees) นั้นคล้ายกับ C4.5 มาก แต่ต่างกันตรงที่รองรับตัวแปรเป้าหมายที่เป็นตัวเลข Regression และไม่คำนวณชุดกฎ CART สร้างต้นไม้ไบนารีโดยใช้คุณลักษณะและเกณฑ์ที่ให้ข้อมูลที่ใหญ่ที่สุดที่ได้รับในแต่ละโหนด Scikit-learn ใช้ Algorithms CART

การจำแนกหมวดหมู่ข้อมูลด้วย Decision Trees ได้รับความนิยมเป็นอย่างมาก และถูกประยุกต์ใช้ในหลากหลายงาน อาทิ เช่น การผลิตและการใช้ยา (Medicine) การผลิตสินค้า (Manufacturing and production) การวิเคราะห์ทางการเงิน (Financial analysis) ดาราศาสตร์ (Astronomy) อณูชีววิทยา (Molecular biology) เป็นต้น สาเหตุที่ Decision Trees ได้รับความนิยมอันเนื่องมากจากเหตุผลหลายประการด้วยกัน เช่น 1) ไม่ต้องการกำหนดค่าพารามิเตอร์ใดๆ เพื่อที่จะทำการสร้าง Decision Trees 2) สามารถจัดการกับข้อมูลที่มีหลายมิติหรือข้อมูลที่มีหลาย Attributes ได้ 3) ผลลัพธ์ที่ได้จะเป็น Trees ที่อยู่ในรูปแบบที่เข้าใจง่าย 4) ขั้นตอนการสร้าง Decision Trees ค่อนข้างง่าย และสามารถทำงานได้อย่างรวดเร็ว 5) มักจะให้ผลการจำแนกข้อมูลที่มีความถูกต้องค่อนข้างสูง แต่อาจขึ้นอยู่กับคุณลักษณะของข้อมูลที่เราใช้ในการสร้างโมเดล

2.3.2 โครงข่ายประสาทเทียม (Artificial Neural Network)

เป็นการจำลองระบบประสาทของสิ่งมีชีวิตขึ้นมา ให้อยู่บนสมการคณิตศาสตร์ หรือโมเดลที่คอมพิวเตอร์สามารถจำลองและคำนวณออกมาได้ Algorithms จะทำการเรียนรู้โดยใช้เครือข่ายประสาทเทียมหลายลำดับชั้น ประกอบไปด้วย 1 ลำดับชั้นสำหรับข้อมูลนำเข้า (Input layer) 1 ลำดับชั้นหรือมากกว่านั้นสำหรับ Hidden layer และ 1 ลำดับชั้นสำหรับข้อมูลออก (Output layer) โดยจะสังเกตได้ว่าแต่ละลำดับชั้นจะประกอบไปด้วยโหนดต่างๆ โดยที่อินพุตสำหรับเครือข่ายจะสอดคล้องกับ Attributes ต่างๆ ของRecord จากชุดข้อมูลสอน ข้อมูลนำเข้าจะถูก

ป้อนเข้าไปยังโหนดต่างๆใน Input layer จากนั้นข้อมูลนำเข้า เหล่านั้นจะผ่าน Input layer และค่าต่างๆจะเดินทางไปสู่ลำดับชั้นต่อไปคือ Hidden layer โดยที่ Output ของ Hidden layer ลำดับชั้นอาจเป็น Input ของ Hidden layer ลำดับชั้นก็เป็นได้ โดยจำนวนลำดับชั้นของ Hidden layer จะมีจำนวนเท่าไรก็ได้แต่โดยทั่วไปมักใช้เพียงแค่ลำดับชั้นเดียว เมื่อผ่านลำดับชั้นที่เป็น Hidden layer แล้วเราจะได้ค่าน้ำหนักที่จะใช้เป็นอิ Input สำหรับ Output layer ที่จะคืนค่าผลของการจำแนกหรือทำนายหมวดหมู่ของข้อมูลสำหรับ Record ที่เป็น Input โดยเมื่อเราพิจารณาที่โหนดต่างๆใน Output layer เราจะเห็นว่า แต่ละโหนดจะทำการรวมค่าน้ำหนักทั้งหมดจากลำดับชั้นก่อนหน้า แล้วทำการประยุกต์ใช้ Nonlinear (Activation) Function เพื่อทำการกำหนดผล

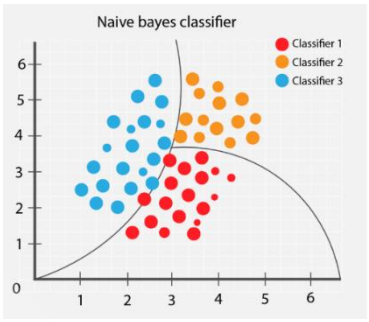


ภาพที่ 2.14 โครงข่ายประสาท (Neural Network)

จากภาพที่ 2.14 โครงข่ายประสาทเทียมแบบที่มีเลเยอร์ซ้อนอยู่หลายชั้น แต่ละชั้นจะถูกระบุเป็นเวกเตอร์ของส่วนประกอบไบนารี โดยมีขอบระหว่างเวกเตอร์กำหนดเป็นเมทริกซ์ของค่าน้ำหนัก พื้นที่การกำหนดค่าของกราฟได้รับจากการแจกแจง Gibbs ด้วย Ising-spin Hamiltonian

2.3.3 การเรียนรู้แบบเบย์ (naive bayesian)

เบย์มีความสามารถจำแนกข้อมูลโดยการเรียนรู้ปัญหาที่เกิดขึ้น เพื่อนำมาสร้างเงื่อนไขการจำแนกข้อมูลใหม่ ใช้การคำนวณหาความน่าจะเป็นในการทำนายผล ที่สามารถคาดการณ์ผลลัพธ์ได้ ในแนวทางที่งานวิจัยนี้ใช้ การสร้างการเรียนรู้ของเครื่อง ด้วย Naive Bayesian โดย Scikit-learn จะอยู่ในรูปของฟังก์ชัน Gaussian NB Gaussian NB สำหรับการจำแนกประเภทความเป็นไปได้ของคุณสมบัติเป็นไปตามลักษณะของ Bayes Theorem

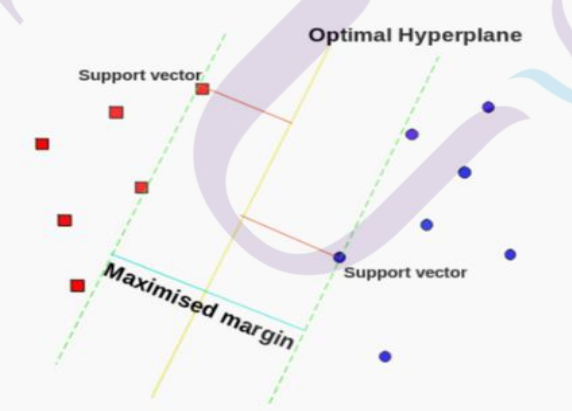


ภาพที่ 2.15 การวิเคราะห์ของทฤษฎีเบย์

จากภาพที่ 2.15 เมื่อนำทฤษฎีของเบย์มาใช้ในงานทางด้าน Data Mining มักจะนิยมนำมาแบ่ง Class ตามการวิเคราะห์จาก Attributes (Attributes) โดยภาพแสดงการแบ่งเป็น 3 กลุ่ม

2.3.4 เทคนิค Support Vector Machine (SVM)

ซัพพอร์ตเวกเตอร์แมชชีน เป็นแบบจำลองเชิงเส้นสำหรับการจำแนกประเภทและปัญหา แนวคิดของ SVM เป็น Algorithms ที่สร้างเส้นหรือไฮเปอร์เพลนซึ่งแยกข้อมูลออกเป็นคลาส ระหว่างข้อมูลของสองคลาส เป้าหมายหลักคืออยากให้เส้นที่แบ่งมีความกว้างมากที่สุด แต่ถ้าเส้นแคบไปก็จะไม่ยอมให้มีข้อมูลอยู่ระหว่างเส้นแบ่ง



ภาพที่ 2.16 เส้นแบ่งกลุ่มข้อมูล Hyper plane (SVM)

จากภาพที่ 2.16 เป็นการแบ่งข้อมูลด้วยเส้นแบ่งกลุ่มข้อมูล Hyper plane สามารถแบ่งได้หลายเส้นแต่จะเลือกเส้นที่มีระยะแบ่งกว้างที่สุด โดยเส้นมีระยะมากที่สุด หาก Margin แคบไปขยับข้อมูลเดียวอาจจะทำให้ข้ามไปอีกฝั่งหนึ่งได้เลยทำให้มีโอกาส Over fit สูง ดังนั้น เราจะเลือก Margin ที่มีค่ามากทำให้เกิด Over fit น้อย หรือเรียกกว่า Soft Margin

2.4 การวิเคราะห์ตัวเลือกคุณลักษณะตามลำดับ (Feature Selector: Backward Selection)

คลาสในโมดูล sklearn.feature_selection สามารถใช้สำหรับการเลือกคุณลักษณะ หรือการลดมิติของชุดข้อมูล เพื่อปรับปรุงคะแนนความแม่นยำ หรือเพื่อเพิ่มประสิทธิภาพในชุดข้อมูล Variance Threshold เป็นแนวทางพื้นฐานในการเลือกคุณสมบัติ โดยจะลบคุณสมบัติทั้งหมดที่มีความแปรปรวนไม่ตรงตามเกณฑ์บางอย่าง โดยค่าเริ่มต้นจะลบคุณลักษณะความแปรปรวนเป็นศูนย์ทั้งหมด เช่น คุณลักษณะที่มีค่าเท่ากันในทุกตัวอย่าง สมมติว่าเรามีชุดข้อมูลที่มีคุณสมบัติ Boolean และเราต้องการลบคุณลักษณะทั้งหมดที่เป็นหนึ่งหรือศูนย์ (on or off) มากกว่า 80% ของตัวอย่าง คุณลักษณะ Boolean คือตัวแปรสุ่ม Bernoulli และความแปรปรวนของตัวแปรดังกล่าว

โดยวิธี Backward Selection เป็นการคัดเลือก Attributes โดยการนำ Attributes ทุกตัวเข้ามาในสมการและดำเนินการพิจารณา Attributes ที่มีค่าสัมประสิทธิ์สหสัมพันธ์บางส่วน (Partial Correlation) กับเกณฑ์ โดยควบคุมอิทธิพลของ Attributes อื่น ๆ ซึ่งมีค่าต่ำที่สุดออกจากสมการแล้วทดสอบว่าค่าที่ลดลงอย่างมีนัยสำคัญทางสถิติหรือไม่ ถ้าพบว่าลดลงอย่างไม่มีนัยสำคัญทางสถิติแสดงว่า Attributes ดังกล่าวไม่ได้ทำให้การทำนายเพิ่มขึ้น แสดงว่าสามารถขจัดออกจากสมการได้เราจึงดำเนินการขจัด Attributes ที่มีความสำคัญน้อยรองลงมาออกไปอีก โดยการขจัด Attributes จนเมื่อพบว่ามีผลทำให้ค่าที่ลดลงอย่างมีนัยสำคัญทางสถิติ ดังนั้น Attributes ดังกล่าวมีความสำคัญต่อการทำนาย หากขจัดตัวแปรดังกล่าวออกจากสมการจะทำให้อำนาจการทำนายลดลง จึงต้องคง Attributes ดังกล่าวไว้ในการสร้างการทำนาย

2.5 การวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis : PCA)

เป็นวิธีการที่ใช้เพื่อสร้างองค์ประกอบใหม่ที่มีขนาดเล็กลงกับข้อมูลที่มีตัวแปรจำนวนมาก โดยวิธีการหาความสัมพันธ์ของตัวแปรเหล่านั้นทำให้เกิดการรวมทำให้ Matrix ที่มีความซับซ้อนเล็กลงหรือที่เรียกว่าทำให้เกิดคุณลักษณะ (Features) ที่มีจำนวนน้อยลง ส่งผลทำให้การสร้างโมเดลในการทำนายใช้เวลาลดลงจึงเป็นที่นิยมในการลดขนาดในการวิเตราะห์ โดยในกรณีที่นิยมทำคือ เมื่อเรามีข้อมูลจำนวนมากและหากเราต้องการใช้ที่ต้องใช้ทรัพยากรในการประมวลผลมาก เช่น Deep learning หากเรามี Features จำนวนมากอาจทำให้เวลาในการประมวลผลใช้เวลามาก

ไม่เหมาะสมกับการใช้งานจริง สมมุติว่าเรามี 300 Features และเราก็ไม่ได้ต้องการรายละเอียดของตัวแปรทั้งหมดนี้ เราก็สามารถสร้างเป็น Components ใหม่ได้ด้วยการกำหนดจาก 300 Features สร้างใหม่อาจเหลือ 4 Components โดยการสร้างตัว Object ขึ้นมาแล้วกำหนดจำนวน Components ในที่นี้คือ x_t หมายถึง ข้อมูลที่ Transform มาแล้ว หลังจากนั้น PCA ก็จะส่งค่ากลับมาใน Scikit - learn เรียกว่า Explained variance หรือ Eigenvector ในการสร้าง Components ใหม่ครั้งนี้ได้ทำให้ได้จำนวน Components ที่ต้องการ

```
pca = PCA(n_components=X_t.shape[1])
# pca = PCA(n_components)

pca.fit_transform(X_t)
print(f'explained_variance (n_components={pca.n_components}) = {pca.explained_variance_}') # Eigenv
print(f'explained_variance_ratio (n_components={pca.n_components}) = {pca.explained_variance_ratio_
print(f'sum explained_variance_ratio = {np.sum(pca.explained_variance_ratio_)}')
```

ภาพที่ 2.17 ตัวอย่างกระบวนการทำ PCA

จากภาพที่ 2.17 แสดงการลดจำนวนข้อมูล 12 Attributes โดยสร้าง 4 Components

2.6 ซอฟต์แวร์และโปรแกรมที่ใช้ในการศึกษา

2.6.1 Anaconda

เดิมชื่อ Continuum Analytics เป็น Distribution เครื่องมือจัดการแพคเกจ Python ในวงการ Data science ช่วยจัดการสภาพแวดล้อมการแจกจ่ายข้อมูลในหลายภาษาคอมพิวเตอร์ เช่น แพ็คเกจภาษา Python / R data science และชุดของแพ็คเกจ Open source สมากกว่า 7,500 รายการ Anaconda เป็น Open source ที่ง่ายต่อการติดตั้งและการใช้งานรวมทั้งยังมีการสนับสนุนจากหลากหลายแหล่งข้อมูลที่ไม่มีค่าใช้จ่ายก่อตั้ง ปี 2012 โดย Travis Oliphant และ Peter Wang ภายใต้ชื่อ Anaconda & Conda และเริ่มให้บริการเซิร์ฟเวอร์อนาคอนดาครั้งแรก นอกจากนี้ยังมีองค์กรที่เกี่ยวข้องจากปี 2012 ถึง 2014 โดย Travis Oliphant เริ่มก่อตั้งและรับสถานะองค์กรการกุศลสาธารณะ โดยไม่แสวงหาผลกำไรในสหรัฐอเมริกาภายใต้ชื่อ PyData, NumFOCUS และเปิด Numba, Bokeh และ Blaze มีความน่าสนใจของแพ็คเกจที่มีอยู่ใน ดังนี้

- มี 24,684 แพ็คเกจใหม่ใน anaconda.org ในปี 2019 สามารถติดตั้งทีละรายการจากที่เก็บ Anaconda ด้วยคำสั่ง conda install และสามารถเพิ่มแพ็กเกจได้ที่ Anaconda.org

- สามารถดาวน์โหลดแพ็คเกจอื่น ๆ โดยใช้คำสั่งที่ติดตั้งกับ Anaconda แพ็คเกจ Pip

- สามารถสร้างแพ็คเกจของตนเองโดยใช้คำสั่งและสามารถแบ่งปันกับผู้อื่นได้โดยการอัปโหลดไปยังAnaconda.org , PyPI หรือที่เก็บอื่นๆ เช่น conda build

ภาพที่ 2.18 แสดงรูปแบบของการรองรับการติดตั้งที่หลากหลายของระบบปฏิบัติการทั้ง Window, MacOS, Linux

ที่มา: https://www.anaconda.com/

2.6.2 Jupyter Notebook

Jupyter คือ Integrated Development Environment นิยมมากในด้านของ Data Science ซึ่งงานด้านนี้ต้องทำงานที่เกี่ยวกับการจัดการข้อมูลเป็นจำนวนมากๆ แล้วยังต้องรายงาน งานวิจัยที่วิจัยไว้ ซึ่งตัว Jupyter Notebook ก็ได้ออกแบบมาตรงตามจุดประสงค์ของการใช้งานไม่ว่าจะเป็นการเรียกใช้งาน library พร้อมทั้งเขียน code และดูผลได้เลย Jupyter Notebook นั้นถูกออกแบบมาให้ทำงานและอ่านได้ง่ายกว่าการใช้งานโปรแกรมแบบปกติ นอกจากนี้ Jupyter ได้พัฒนาในแบบเปิดกับผู้ให้บริการของเว็บโฮสติ้งสำหรับการพัฒนาซอฟต์แวร์ Open source GitHub โดยผ่านความเห็นชอบของ Jupyter community

Project Jupyter เป็นโครงการ Open source ที่ไม่แสวงหาผลกำไรเกิดจากโครงการ IPython และในปี 2014 โครงการ Jupyter ถูกแยกออกจาก IPython จุดเริ่มต้นจาก Fernando Pérez โดย Project Jupyter เป็นการอ้างอิงถึงการเขียนภาษาโปรแกรมหลักสามภาษาที่รองรับโดย Jupyter ได้แก่ Julia, Python และ R และยังเป็นการแสดงความเคารพต่อสมุดบันทึกของกาลิเลโอที่บันทึกการค้นพบดวงจันทร์ของดาวพฤหัสบดี Project Jupyter ได้พัฒนาและสนับสนุนผลิตภัณฑ์

คอมพิวเตอร์เชิงโต้ตอบ Jupyter Notebook, JupyterHub และ Jupyter Lab เป็นซอฟต์แวร์ Open source ภายใต้เงื่อนไขเสรีของใบอนุญาต BSD

ภาพที่ 2.19 มาตรฐานการประมวลผลเชิงโต้ตอบในภาษาโปรแกรม

จากภาพที่ 2.19 Project Jupyter มีขึ้นเพื่อพัฒนาซอฟต์แวร์ Open source สำหรับการประมวลผลเชิงโต้ตอบเพื่องานวิเคราะห์ด้านสถิติ ในภาษาโปรแกรม เช่น Julia, R, Python เป็นต้น

2.6.3 Library

ไลบรารี่ (library) เปรียบเสมือนโปรแกรมสำเร็จรูปที่เก็บฟังก์ชันการทำงานที่เฉพาะทาง ส่วนที่รวบรวมกระบวนการ (process) และฟังก์ชันย่อย (subroutine) ต่าง ๆ ซึ่งอาจจะรวมซอร์สโค้ดที่จำเป็นต้องใช้ในการเขียนซอฟต์แวร์หรือใช้ในการทำงานของโปรแกรม ประโยชน์ของไลบรารี่สำเร็จรูปที่มากับโปรแกรมก็คือ ผู้ใช้งานไม่จำเป็นที่จะต้องสร้างไลบรารี่ขึ้นมาใหม่เองทั้งหมด แต่สามารถนำไลบรารี่ที่ได้มีการพัฒนาไว้อยู่แล้วมาใช้งานได้เลย เช่น ฟังก์ชัน print ใน Matplotlib หรือการอ่านไฟล์เพื่อช่วยเตรียมข้อมูลให้สมบูรณ์ก่อนที่จะนำไปคำนวณใน Pandas

2.6.4 Scikit-Learn

Scikit-Learn เป็น Bundle library แบบ Open source ชนิด BSD License ใช้ภาษา Python, Numpy และ Scipy มีจุดเด่น คือฟังก์ชันในการแบ่งประเภทข้อมูล การแบ่งกลุ่มข้อมูลการวิเคราะห์หลากหลาย ที่รวมAlgorithms Machine Learningที่เหมาะกับทั้งผู้เชี่ยวชาญและผู้เริ่มต้นใช้งานการเรียนรู้ของเครื่อง Scikit-Learn เป็นโครงการที่เริ่มต้นในปี 2007 ภายใต้ชื่อโครงการ Google Summer of Code มี David Cournapeau ดูแล ต่อมา Matthieu Brucher ก็เริ่มเข้ามาทำงานในโครงการนี้ เขาสนใจ และนำมาเป็นส่วนหนึ่งของวิทยานิพนธ์ของเขาในปี 2010 ต่อมา Fabian Pedregosa,Gael Varoquaux, Alexandre Gramfort และVincent Michel จากสถาบันวิจัยวิทยาศาสตร์ดิจิทัลและเทคโนโลยีแห่งชาติ INRIA ได้เป็นผู้นำของโครงการต่อ และได้เผยแพร่สู่สาธารณะเป็นครั้งแรกในวันที่ 1 กุมภาพันธ์ 2010 ตั้งแต่นั้นมาก็มีการเผยแพร่ผลงานของโครงการออกมาอย่าง

ต่อเนื่อง และทำให้เกิดการเติบโตของประชาคมระหว่างประเทศในการพัฒนา Scikit-Learn เป็นโครงการที่ขับเคลื่อนโดยชุมชนและโครงการนี้ได้รับทุนจากรัฐ สถาบันเอกชนจึงทำให้มั่นใจได้ถึงความยั่งยืนของโครงการ นอกจากนี้ยังได้รับการช่วยเหลือจาก Microsoft Azure, Travis Cl, Circle Cl, Anaconda Inc. สำหรับการประมวลผล และการจัดเก็บข้อมูลบนเซิร์ฟเวอร์โดยการผสานรวมแบบต่อเนื่อง

ในปัจจุบันมีบริษัทและองค์กรจำนวนมากใช้บริการของ Scikit-Learn ได้แก่ J.P.Morgan, Spotify, INRIA, Booking.com, Change.org, เป็นต้น โดยได้ใช้งาน Scikit-Learn ดังนี้

Stephen Simmons, VP, Athena Research, JPMorganได้กล่าวว่า ใช้ Scikit-learn เป็นส่วนของชุดเครื่องมือ Machine Learning โดยใช้ Python JPMorgan ใช้งานในทุกส่วนของธนาคารสำหรับการจัดหมวดหมู่การวิเคราะห์เชิงคาดการณ์และ Machine Learning อื่น ๆ อีกมากมาย ใช้ในส่วน API เพราะสามารถใช้กับAlgorithmsที่หลากหลายและมีคุณภาพ Scikit-learn เข้าถึงใช้งานได้ง่ายและมีประสิทธิภาพมาก

Erik Bernhardsson, Engineering Manager Music Discovery & Machine Learning, Spotify กล่าวว่าSpotify ใช้ Scikit-learn เนื่องจากมีชุดเครื่องมือที่มีการใช้งานโมเดลที่ทันสมัยมากมายและทำให้ง่ายต่อการใช้ร่วมกับกับแอพพลิเคชั่นที่มีอยู่ เราใช้มันค่อนข้างมากในการแนะนำเพลงที่ Spotify และเราคิดว่ามันเป็นแพ็คเกจ Machine Learning ที่ออกแบบมาอย่างดีที่สุด

INRIA ใช้ scikit-learn เพื่อสนับสนุนการวิจัยขั้นพื้นฐานระดับแนวหน้าในหลาย ๆ ทีม: Parietal for neuroimaging, Lear สำหรับการมองเห็นด้วยคอมพิวเตอร์, Visages สำหรับการวิเคราะห์ภาพทางการแพทย์, Privatics for security โครงการนี้เป็นเครื่องมือที่ยอดเยี่ยมในการจัดการกับการใช้งาน Machine Learning เนื่องจากมีประสิทธิภาพและมีความหลากหลาย และใช้งานง่าย มีเครื่องมือที่ดีเหมาะสำหรับนักเรียนที่กำลังศึกษา

Booking.com ใช้Algorithms Machine Learning สำหรับแอปพลิเคชันต่างๆ เช่น แนะนำโรงแรมและจุดหมายปลายทางให้กับลูกค้า ตรวจจับการจองที่หลอกลวง หรือกำหนดเวลาตัวแทนบริการลูกค้าของเรา Scikit-learn เป็นเครื่องมือที่เราใช้Algorithmsช่วยสำหรับงานทำนาย API และทำให้ใช้งานง่าย นักพัฒนา scikit-learn ทำงานได้อย่างดีในการผสมผสานการใช้งานที่ทันสมัยและAlgorithmsใหม่เข้ากับแพ็คเกจ ดังนั้น scikit-learn จึงช่วยให้สามารถเข้าถึงAlgorithmsที่หลากหลายได้อย่างสะดวกและช่วยให้เราค้นหาเครื่องมือที่เหมาะสมสำหรับงานที่เหมาะสมได้อย่างดี

Change.org ใช้ Random Forest Classifier ของ scikit-learn โดยสามารถให้ Machine Learning ทำงานอัตโนมัติในระบบการของเรา เพื่อผลักดันการกำหนดเป้าหมายในการโต้ตอบอีเมล

ที่เข้าถึงผู้ใช้หลายล้านคนทั่วโลกในแต่ละสัปดาห์ มีความสะดวกในการใช้งานประสิทธิภาพและ Algorithms ที่หลากหลายของ scikit-learn ที่ Change.org นำมาใช้นั้นได้พิสูจน์แล้วว่ามีประสิทธิภาพในการจัดหาแหล่งข้อมูลที่เชื่อถือสูงเพื่อตอบสนองความต้องการใช้งาน Machine Learning ของเรา

2.6.5 Pandas

เป็น library การจัดการข้อมูล (Data Manipulation) ช่วยจัดการข้อมูลที่จะต้อง import และแสดงข้อมูลออก โดยใช้คำสั่งจาก Pandas และสามารถอ่านไฟล์ได้เกือบทั้งหมดไม่ว่าจะเป็นแบบ FWF,CSV, Excel, Zip, TSV นอกจากนี้ยังมีคำสั่งที่จะช่วยเตรียมข้อมูลให้สมบูรณ์ก่อนที่จะนำไปคำนวณอีกด้วย จึงเป็น Library ที่นิยมใช้ในด้าน Data Science

การพัฒนาแพนด้าเริ่มขึ้นที่ AQR Capital Management ในปี 2008 เปิดให้บริการ Open source และได้รับการสนับสนุนจาก Community โดยมีเป้าหมายที่จะเป็นส่วนประกอบพื้นฐานระดับสูงสำหรับการวิเคราะห์ข้อมูลจริงที่ใช้งานได้จริงใน Python นอกจากนี้ยังมีเป้าหมายที่กว้างขึ้นในการเป็นเครื่องมือวิเคราะห์ / จัดการข้อมูล Open source ที่มีประสิทธิภาพและยืดหยุ่นที่สุดที่มีอยู่ในทุกภาษา โดยมีคุณสมบัติดังนี้

- สามารถการจัดการข้อมูลด้วยการจัดทำดัชนีในตัวออบเจ็กต์ Data Frame ที่รวดเร็วและมีประสิทธิภาพสำหรับ

- เครื่องมือสำหรับการอ่านและเขียนข้อมูลระหว่างโครงสร้างข้อมูลในหน่วยความจำและรูปแบบต่างๆ เช่น ไฟล์ CSV และข้อความ, Microsoft Excel, ฐานข้อมูล SQL และรูปแบบ HDF5 ที่รวดเร็ว

- มีความสามารถในการจัดตำแหน่งข้อมูลที่ดี และการจัดการข้อมูลที่ขาดหายไปในตัวรองรับการจัดตำแหน่งตามLabelsอัตโนมัติในการคำนวณและจัดการข้อมูลที่ยุ่งเหยิงให้เป็นรูปแบบที่เป็นระเบียบได้

- การแบ่งส่วนตาม label การจัดทำดัชนีแบบ fancy index และการย่อยชุดข้อมูลขนาดใหญ่ รวมทั้งสามารถแทรกคอลัมน์ และลบออกจากโครงสร้างข้อมูลเพื่อรองรับความไม่แน่นอนของ size mutability

- ฟังก์ชันอนุกรมเวลา การสร้างช่วงวันที่และการแปลงความถี่สถิติการย้าย window statistics การเปลี่ยนวันที่และการ time offsets แม้แต่สร้างการชดเชยเวลาเฉพาะโดเมนและเข้าร่วมอนุกรมเวลาโดยไม่สูญเสียข้อมูล

ตั้งแต่ปี 2558 แพนด้าเป็นโครงการที่ได้รับการสนับสนุนจาก Num FOCUS จึงช่วยให้การพัฒนาแพนด้าประสบความสำเร็จในฐานะโครงการ Open source สระดับโลก และPython ที่มี

แพนด้าถูกนำไปใช้ในโดเมนทางวิชาการและการค้าที่หลากหลายรวมถึงการเงิน เศรษฐศาสตร์ สถิติ การโฆษณา การวิเคราะห์เว็บมากมาย

2.6.6 Numpy

library คณิตศาสตร์และสถิติ (Mathematics and Statistics) ที่เกี่ยวกับฟังก์ชั่นทางคณิตศาสตร์และการคำนวณต่างๆ โดยจะเข้าไปจัดการข้อมูลที่เป็นตารางแบบ array โดยเป็นโครงการ Open source ที่มีวัตถุประสงค์เพื่อเปิดใช้งานการคำนวณเชิงตัวเลขด้วย Python สร้างขึ้นในปี 2005 โดยสร้างขึ้นช่วงแรกของไลบรารีชื่อ Numeric และ Numarray Numpy จะเป็นซอฟต์แวร์ Open source เผยแพร่ภายใต้เงื่อนไขเสรีของใบอนุญาต BSD ต่อมา Numpy ได้รับการพัฒนาบน GitHub ผ่านความเห็นพ้องของ Numpy communityและ Python community

2.6.7 Matplotlib

เป็น library การแสดงผลข้อมูล (Data Visualization) พื้นฐานในการสร้างกราฟจาก array กรอกค่าแสดงความสัมพันธ์ของตัวแปร ซึ่งสามารถสร้างกราฟในแบบ 2 มิติ ใน Python แม้ว่าจะมีต้นกำเนิดในการเลียนแบบคำสั่งกราฟิก MATLAB แต่ก็ไม่ได้อยู่ภายใต้พื้นฐาน MATLAB และสามารถใช้ในรูปแบบ Pythonic เชิงวัตถุได้ แม้ว่า Matplotlib จะเขียนด้วย Python แต่ก็ใช้ NumPy และโค้ดส่วนขยายอื่น ๆ หลายส่วน เพื่อให้ได้ประสิทธิภาพที่ดีรวมทั้งมีความสามารถในอาร์เรย์ขนาดใหญ่

การออกแบบ Matplotlib บนพื้นฐานที่ว่าสามารถสร้างพล็อตง่าย ๆ ด้วยคำสั่งเพียงไม่กี่คำสั่งหรือเพียงคำสั่งเดียว หากต้องการดูฮิสโตแกรมของข้อมูล ก็ไม่จำเป็นต้องสร้างอินสแตนซ์ออบเจ็กต์วิธีการเรียกการตั้งค่าคุณสมบัติและอื่น ๆ ก็ใช้งานได้ Code Matplotlib แบ่งตามแนวคิดออกเป็นสามส่วน

1) pylab interface คือ ชุดของฟังก์ชันที่จัดทำโดย pylab ซึ่งอนุญาตให้ผู้ใช้สร้างพล็อตด้วยโค้ดที่ค่อนข้างคล้ายกับโค้ดสร้างรูป MATLAB

2) Pyplot tutorial คือ Matplotlib frontend หรือ Matplotlib API เป็นชุดของคลาสที่ทำหน้าที่ในการ heavy lifting สร้างและจัดการตัวเลขข้อความเส้นพล็อตและอื่น ๆ

3) Artist tutorial คือ interface นามธรรมที่ไม่รู้อะไรเกี่ยวกับเอาต์พุต แบ็กเอนด์เป็นอุปกรณ์วาดภาพขึ้นอยู่กับอุปกรณ์ หรือที่เรียกว่าตัวแสดงผลที่เปลี่ยนการแสดงส่วนหน้าเป็นสำเนาหรืออุปกรณ์แสดงผล ว่าตัวแบ็กเอนด์คืออะไร เช่น แบ็กเอนด์ PS สร้าง PostScript® hardcopy, SVG สร้างสำเนากราฟิกแบบเวกเตอร์ที่ปรับขนาดได้, Agg สร้างเอาต์พุต PNG โดยใช้ไลบรารี Anti-Grain Geometry คุณภาพสูงที่มาพร้อมกับ Matplotlib, GTK ฝัง Matplotlib ในแอปพลิเคชัน

Gtk +, GTKAgg ใช้ตัวแสดงผล Anti-Grain เพื่อ สร้างรูปและฝังไว้ในแอปพลิเคชัน Gtk + และอื่น ๆ สำหรับ PDF, WxWidgets, Tkinter และอื่น ๆ

การประยุกต์ใช้ Matplotlib ถูกใช้โดยคนทั่วโลกจำนวนมากในบริบทต่างๆ ดังนี้

- ต้องการสร้างไฟล์ PostScript โดยอัตโนมัติเพื่อส่งไปยังเครื่องพิมพ์หรือผู้เผยแพร่ อื่น ๆ โดยสามารถปรับใช้ Matplotlib บนเว็บแอ็พพลิเคชันเซิร์ฟเวอร์เพื่อสร้างเอาต์พุต PNG สำหรับรวมไว้ในหน้าเว็บที่สร้างแบบไดนามิก

- ใช้ Matplotlib แบบโต้ตอบจาก Python shell ใน Tkinter บน Windows

- การฝัง Matplotlib ในแอปพลิเคชัน Gtk + EEG ที่ทำงานบน Windows, Linux และ Macintosh OS X

2.7 เอกสารงานวิจัยที่เกี่ยวข้อง

นนท์ บุญนิธิประเสริฐ และชัยพร เขมะภาตะพันธ์ ที่ได้นำเสนอวิธีการกรองสแปมในระบบส่งข้อความของโทรศัพท์เคลื่อนที่ ด้วยวิธีการกรองโดยใช้ Support Vector Machine และ Naive Bayesianหลักการของการกรองสแปม ในงานวิจัยจะมีการแบ่งคลาสเป็น 2 แบบ คือใช่และไม่ใช่ ใช้ลักษณะเป็นเหมือนกันกับการเลือกระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์ซึ่งแทนด้วยเลข 1 หรือแผนศิลป์แทนด้วยเลข 2 โดยผลทดสอบด้วยวิธีการกรองแบบ Support Vector Machine มีความถูกต้องในการกรองข้อมูลสูงกว่า Naive Bayesian แต่วิธีการกรองแบบ Naive Bayesian ใช้เวลาในการประมวลผลน้อยกว่า

อรทิพย์ เลื่อยงาม และชัยพร เขมะภาตะพันธ์ วิจัยเรื่อง การจัดประเภทเอกสารด้วยวิธีเอสวีเอ็ม เพื่อการป้องกันเอกสารรั่วไหล เป็นการออกแบบระบบการแยกประเภทเอกสารสำคัญออกจากเอกสารทั่วไปเพื่อนำไปเป็นอินพุตของระบบการป้องกันข้อมูลรั่วไหล DLP พบว่าการแยกประเภทเอกสาร โดยใช้ข้อมูลชุดฝึกสอน และข้อมูลชุดทดสอบทั้งสองประเภท คือเอกสารความลับและเอกสารทั่วไป สามารถแยกประเภทเอกสารตามที่ได้กำหนดประเภทเอกสารไว้ ซึ่งได้ผลเป็นที่น่าพอใจ

สุพัฒน์กุล ภัคโชค ศึกษาปัจจัยที่มีผลต่อการเลือกแผนการเรียน และ ความสามารถในการศึกษาในแผนการเรียน ด้วยเทคนิคเหมืองข้อมูล โดยใช้ผลการเรียนในรายวิชาต่างๆ และข้อมูลแบบสอบถามวิธีการ ตัดสินใจเลือกแผนการเรียนของนักเรียนระดับมัธยมศึกษาตอนปลาย โดยมีจำนวนตัวอย่างจาก นักเรียนระดับมัธยมศึกษาปลาย 850 คน ผลการวิจัยที่ได้แสดงให้เห็นว่าตัวแบบการเลือกแผนการเรียนระดับมัธยมศึกษาตอนปลายสามารถ บ่งบอกได้ว่าปัจจัยใดมีผลต่อ

การตัดสินใจเลือกแผนการเรียนของนักเรียน และให้ค่าความถูกต้อง ในการแนะนำแผนการเรียน ร้อยละ 79.76 จากตัวแบบดังกล่าวทำให้ทราบว่าวิชาพื้นฐานในระดับ มัธยมศึกษาตอนต้น ได้แก่ วิชาภาษาไทย คณิตศาสตร์ วิทยาศาสตร์ สังคมศึกษา ศาสนาและ วัฒนธรรม และภาษาอังกฤษ รวมทั้งผลการเรียนเฉลี่ยสะสมในระดับมัธยมศึกษาตอนต้นเป็นปัจจัย หลักที่มีผลโดยตรงต่อการตัดสินใจเลือกแผนการเรียนของนักเรียนและสามารถศึกษาในแผนการเรียนนั้นได้อย่างประสบความสำเร็จ

วุฒิชัย กำจรกิตติคุณ และชัยพร เขมะภาตะพันธ์ วิจัยเรื่อง การวิเคราะห์ความผิดปกติข้อมูลการใช้ไฟฟ้าด้วยนาอีฟเบย์ เป็นการคัดกรองความผิดปกติของข้อมูลการใช้ไฟฟ้ารายใหญ่ โดยนำสถิติการใช้ไฟฟ้ามาวิเคราะห์รูปแบบการใช้กำหลังไฟฟ้ารายเดือนควบคู่กับความสัมพันธ์ของพลังงานที่ใช้ในแต่ละเฟส พบว่า จำนวนครั้งในการตรวจสอบที่ไม่จำเป็นลดลงเมื่อเทียบกับการดำเนินการแบบเดิมที่ต้องตรวจสอบทุกกรณี และสร้างความน่าเชื่อถือให้กับระบบไฟฟ้าอีกด้วย

อตินติ วัฒนบุรานนท์ เสนอวิธีการจำแนกประโยคของบทวิจารณ์เกมแอ็คชัน โดยแบ่งเป็น 5 ขั้นตอนคือ 1) การระบุและนิยามปัจจัยที่ ส่งผลต่อคุณภาพของเกมและปัญหาของเกม 2) การสร้างคลังคำศัพท์ของปัจจัยที่นิยามไว้ด้วย การ สกัดคำอธิบายการระบุและนิยามปัจจัยการสกัดจากบทวิจารณ์เกม และการใช้เครื่องมือเวิร์ดเน็ตใน การขยายคลังคำศัพท์ 3) การวิเคราะห์เพื่อจำแนกประโยคของบทวิจารณ์ทั้งในด้านความคิดเห็นเชิง บวกและเชิงลบ และด้านปัจจัยที่นิยามไว้ 4) การพัฒนาเครื่องมือตามแนวคิดที่นำเสนอ 5) การ ประเมินประสิทธิผลของเครื่องมือด้วยการทดลองจากการคำนวณค่าความแม่นยำ โดยการตรวจสอบ การจำแนกประโยคของบทวิจารณ์เกมโดยเครื่องมือเปรียบเทียบกับการจำแนกโดยผู้เชี่ยวชาญประสิทธิผลของเครื่องมือในการจำแนกได้เหมือนกันทั้งหมดกับผู้เชี่ยวชาญในรูปของค่า ความแม่นยำเฉลี่ย ของการจำแนกความคิดเห็นเชิงบวกและเชิงลบมีค่าร้อยละ 59.51 ของการจำแนก ปัจจัยที่ส่งผลต่อคุณภาพของเกมและปัญหาของเกมมีค่าร้อยละ 64.46 และ 81.01 ตามลำดับ

บทที่ 3
การดำเนินงานวิจัย

การดำเนินงานวิจัยเรื่องแนวโน้มการเลือกแผนการเรียนมัธยมศึกษาตอนปลายด้วยการใช้ Machine Learning แบบ Supervised Learning เปรียบเทียบ 4 วิธี ได้แก่ Decision Tree, Artificial Neural Network, Naive Bayes, Support Vector Machine เพื่อหาแนวโน้มการตัดสินใจเลือกแผนการเรียนมัธยมศึกษาตอนปลาย ระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์หรือแผนศิลป์ สำหรับนักเรียนมัธยมศึกษาโรงเรียนสาธิต "พิบูลบำเพ็ญ" มหาวิทยาลัยบูรพา

ขั้นตอนการดำเนินการวิจัย

3.1 การเลือกข้อมูล (Data Selection)

3.2 การกลั่นกรองข้อมูล (Data Cleansing)

3.3 การแปลงข้อมูล (Data Transformation)

3.4 การสร้างแบบจำลอง (Model) ทดสอบและเปรียบเทียบประสิทธิภาพ

3.5 การแปรผลและการประเมินผล (Interpretation Evaluation)

3.1 การเลือกข้อมูล (Data Selection)

ผู้วิจัยศึกษาข้อมูลที่เกี่ยวข้องกับกลุ่มตัวอย่างมัธยมศึกษาตอนต้นมากำหนด Attributes เพื่อการวิเคราะห์หาแนวโน้มการตัดสินใจเลือกแผนการเรียนมัธยมศึกษาตอนปลายระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์ หรือแผนศิลป์ปีการศึกษา 2554 -2556 ได้แก่ เพศ ระดับผลการเรียนรายวิชาจำนวน 6 ภาคการศึกษา เกรดเฉลี่ยการศึกษาและข้อมูลอื่นดังนี้

1) เพศ

2) ภูมิลำเนา

3) อาชีพของผู้ปกครอง

4) รายได้ผู้ปกครอง

5) แผนการเรียน

6) วิชาคณิตศาสตร์ ภาคการศึกษาที่ 1-1, 2-1, 3-1, 3-2, 4-1, 4-2, 5-1, 5-2, 6-1

7) วิชาวิทยาศาสตร์ ภาคการศึกษาที่ 1-1, 2-1, 3-1, 4-1, 5-1, 6-1

8) วิชาภาษาอังกฤษ ภาคการศึกษาที่ 1-1, 2-1, 3-1, 4-1, 5-1, 6-1

9) วิชาภาษาไทย ภาคการศึกษาที่ 1-1, 2-1, 3-1, 4-1, 5-1, 6-1

10) วิชาสังคม ภาคการศึกษาที่ 1-1, 2-1, 3-1, 4-1, 5-1, 6-1

11) วิชาศิลปะ ภาคการศึกษาที่ 1-1, 2-1, 3-1, 4-1, 5-1, 6-1

12) วิชาสุขศึกษา ภาคการศึกษาที่ 1-1, 2-1, 3-1, 4-1, 5-1, 6-1

13) วิชาประวัติศาสตร์ ภาคการศึกษาที่ 1-1, 2-1, 3-1, 4-1, 5-1, 6-1

14) วิชาพลศึกษา ภาคการศึกษาที่ 1-1, 2-1, 3-1, 4-1, 5-1, 6-1

15) วิชาการงานอาชีพและเทคโนโลยี ภาคการศึกษาที่ 1-1, 2-1, 3-1, 4-1, 5-1, 6-1

16) เกรดเฉลี่ยมัธยมศึกษาปีที่ 3

17) เกรดเฉลี่ยมัธยมศึกษาปีที่ 6

คุณลักษณะหรือแอททริบิวต์ดังกล่าวผู้วิจัยได้รวบรวมทุกขอบเขตที่มีอยู่ในระบบข้อมูล ได้จำนวน 70 Attributes

ตารางที่ 3.1 การกำหนดตัวแปร

Attributes	ความหมาย
sex	เพศ
geo	ภูมิลำเนา
career	อาชีพของผู้ปกครอง
path	แผนการเรียน
math	วิชาคณิตศาสตร์
sci	วิชาวิทยาศาสตร์
eng	วิชาภาษาอังกฤษ
thai	วิชาภาษาไทย
soc	วิชาสังคม
art	วิชาศิลปะ
heal	วิชาสุขศึกษา
his	วิชาประวัติศาสตร์
pe	วิชาพลศึกษา
we	วิชาการงานอาชีพและเทคโนโลยี
gpax3	เกรดเฉลี่ยมัธยมศึกษาปีที่ 3
Gpax6	เกรดเฉลี่ยมัธยมศึกษาปีที่ 6

3.2 การกลั่นกรองข้อมูล (Data Cleansing)

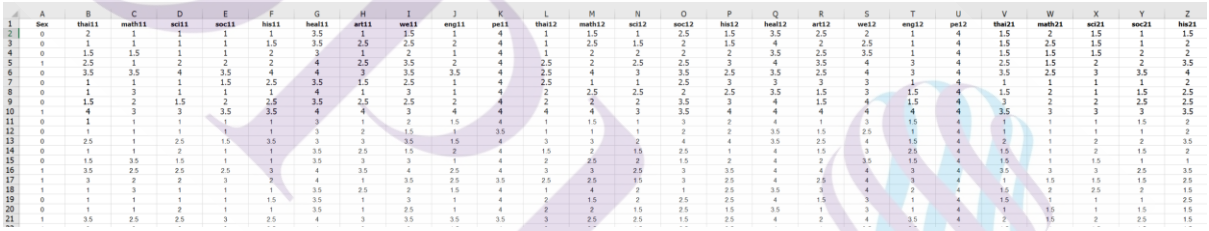
ผู้วิจัยรวบรวมข้อมูลทั้งรูปแบบเอกสารและอิเล็กทรอนิกส์ในระบบฐานข้อมูลรวมก่อนการคัดเลือก 70 แอทริบิวต์ มาผ่านกระบวนการดังนี้

1) นำข้อมูลมาปรับปรุงในส่วนของค่าข้อมูลที่ขาดหาย (Missing Value) กำจัดข้อมูลว่าง (Null Value) โดยใช้วิธีการตัดข้อมูลที่ไม่สมบูรณ์ออก และรวมเข้าเป็นเอกสารอิเล็กทรอนิกส์ไฟล์ Excel เดียวกัน เพื่อตรวจสอบความสมบูรณ์ของข้อมูลแสดงข้อมูลดังตาราง

2) นำข้อมูลมาตรวจความไม่สอดคล้อง (Inconsistent Data) เช่น ข้อมูลรายวิชาในแต่ละภาคการศึกษาของผู้เรียนที่เรียนแผนการเรียนต่างกันใช้ค่าแทนข้อมูลที่ต่างกันมาปรับปรุง

3) นำมาเตรียมข้อมูลโดยการแยกข้อมูลมาจัดกระทำให้อยู่ในรูปแบบจำนวน 6 ภาคการศึกษา และนำค่าเกรดเฉลี่ยตลอดหลักสูตรชั้นมัธยมศึกษาตอนต้นและตอนปลาย มาลงในรูปแบบตารางไฟล์ Excel เดียวกันเพื่อตรวจสอบความสมบูรณ์ของข้อมูล

ตารางที่ 3.2 การรวมเอกสารอิเล็กทรอนิกส์ในไฟล์ Excel



3.3 การแปลงข้อมูล (Data Transformation)

3.3.1 การกำหนด Code ข้อมูล

ผู้วิจัยได้ทำการแปลงข้อมู เริ่มจากการกำหนด Code ข้อมูลและเปลี่ยนรูปข้อมูลให้อยู่ในรูปแบบที่เหมาะสมกับการทำ Machin Learning วิธี Decision Tree, Naive Bayes, Artificial Neural Network และ Support Vector Machine โดยมี Attributes ดังนี้

1) Attributes ที่ใช้ค่าตามความเป็นจริง คือ ข้อมูลระดับผลการเรียนรายวิชา วิชาคณิตศาสตร์, วิชาวิทยาศาสตร์, วิชาภาษาอังกฤษ, วิชาภาษาไทย, วิชาสังคม, วิชาศิลปะ, วิชาสุขศึกษา, วิชาประวัติศาสตร์, วิชาพลศึกษา, วิชาการงานอาชีพและเทคโนโลยี, เกรดเฉลี่ยมัธยมศึกษาปีที่ 3, เกรดเฉลี่ยมัธยมศึกษาปีที่ 6

2) Attributes ให้ค่า เท่ากับ 0 และ 1 (Binary Number) คือ แผนการเรียน เพศ

ตารางที่ 3.3 การแปลงข้อมูลและการกำหนด Code

No.	Attributes	Detail	Code
1	Sex	เพศ	0 = ชาย 1 = หญิง
2	Path	แผนการเรียน *ใช้เป็น Attributes เป้าหมาย	แผนวิทยาศาสตร์และคณิตศาสตร์ SM = 1 แผนศิลป์ Art = 2
3	Official (O)	ข้าราชการ/เจ้าหน้าที่ของรัฐ	1 = ใช่ 0 = ไม่ใช่
4	State Enterprise Employees (SEEM)	พนักงานรัฐวิสาหกิจ	1 = ใช่ 0 = ไม่ใช่
5	Company employee (CEM)	พนักงานบริษัทเอกชน	1 = ใช่ 0 = ไม่ใช่
6	Business (B)	ธุรกิจส่วนตัว/ค้าขาย	1 = ใช่ 0 = ไม่ใช่
7	Employee (EM)	รับจ้าง/ลูกจ้าง	1 = ใช่ 0 = ไม่ใช่
8	Farmer (F)	เกษตรกรรม/ปศุสัตว์/ประมง	1 = ใช่ 0 = ไม่ใช่
9	N	ภาคเหนือ	1 = ใช่ 0 = ไม่ใช่
10	NE	ภาคตะวันออกเฉียงเหนือ	1 = ใช่ 0 = ไม่ใช่
11	E	ภาคตะวันออก	1 = ใช่ 0 = ไม่ใช่
12	S	ภาคใต้	1 = ใช่ 0 = ไม่ใช่

ตารางที่ 3.3 การแปลงข้อมูลและการกำหนด Code (ต่อ)

No.	Attributes	Detail	Code
13	W	ภาคตะวันตก	1 = ใช่ 0 = ไม่ใช่
14	C	ภาคกลาง	1 = ใช่ 0 = ไม่ใช่
15	35K	รายได้ผู้ปกครอง น้อยกว่า 35,000	ใช้ค่าตามจริง
16	35K-50K	รายได้ผู้ปกครอง 35,001 – 50,000	ใช้ค่าตามจริง
17	50K-65K	รายได้ผู้ปกครอง 50,001 - 65,000	ใช้ค่าตามจริง
18	65K-80K	รายได้ผู้ปกครอง 65,001 - 80,000	ใช้ค่าตามจริง
19	80K-95K	รายได้ผู้ปกครอง 80,001 - 95,000	ใช้ค่าตามจริง
20	95K-110K	รายได้ผู้ปกครอง 95,001 - 110,000	ใช้ค่าตามจริง
21	110k	รายได้ผู้ปกครอง 110,001 ขึ้นไป	ใช้ค่าตามจริง
22	Math1-1, Math2-1, Math3-1, Math3-2, Math4-1, Math4-2, Math5-1, Math5-2, Math6-1	วิชาคณิตศาสตร์ ผลการเรียน 6 ภาคการศึกษา	ใช้ค่าผลการเรียนตามจริง
23	Sci1-1, Sci1-2, Sci2-1, Sci2-2, Sci3-1, Sci3-2	วิชาวิทยาศาสตร์ ผลการเรียน 6 ภาคการศึกษา	ใช้ค่าผลการเรียนตามจริง
24	Eng1-1, Eng1-2, Eng2-1, Eng2-2, Eng3-1, Eng3-2	วิชาภาษาอังกฤษ ผลการเรียน 6 ภาคการศึกษา	ใช้ค่าผลการเรียนตามจริง

ตารางที่ 3.3 การแปลงข้อมูลและการกำหนด Code (ต่อ)

No.	Attributes	Detail	Code
25	Thai1-1, Thai1-2, Thai2-1, Thai2-2, Thai3-1, Thai3-2	วิชาภาษาไทย ผลการเรียน 6 ภาคการศึกษา	ใช้ค่าผลการเรียนตามจริง
26	Soc1-1, Soc1-2, Soc2-1, Soc2-2, Soc3-1, Soc3-2	วิชาสังคม ผลการเรียน 6 ภาคการศึกษา	ใช้ค่าผลการเรียนตามจริง
27	Art1-1, Art1-2, Art2-1, Art2-2, Art3-1, Art3-2	วิชาศิลปะ ผลการเรียน 6 ภาคการศึกษา	ใช้ค่าผลการเรียนตามจริง
28	Heal1-1, Heal1-2, Heal2-1, Heal2-2, Heal3-1, Heal3-2	วิชาสุขศึกษา ผลการเรียน 6 ภาคการศึกษา	ใช้ค่าผลการเรียนตามจริง
29	His1-1, His1-2, His2-1, His2-2, His3-1, His3-2	วิชาประวัติศาสตร์ ผลการเรียน 6 ภาคการศึกษา	ใช้ค่าผลการเรียนตามจริง
30	Pe1-1, Pe1-2, Pe2-1, Pe2-2, Pe3-1, Pe3-2	วิชาพลศึกษา ผลการเรียน 6 ภาคการศึกษา	ใช้ค่าผลการเรียนตามจริง
31	We1-1, We1-2, We2-1, We2-2, We3-1, We3-2	วิชาการงานอาชีพและเทคโนโลยี ผลการเรียน 6 ภาคการศึกษา	ใช้ค่าผลการเรียนตามจริง
32	gpax3	เกรดเฉลี่ยมัธยมศึกษาปีที่ 3	ใช้ค่าผลการเรียนตามจริง
33	Gpax6	เกรดเฉลี่ยมัธยมศึกษาปีที่ 6	ใช้ค่าผลการเรียนตามจริง

จากตารางที่ 3.3 แสดงรูปแบบของข้อมูลในแต่ละ Attributes 2 ประเภทคือ 1) ข้อมูลที่มี 2 ค่า (Binary Number) คือ 0 กับ 1 เช่น แผนการเรียน และเพศ 2) ข้อมูลที่เป็นจำนวนจริง เช่น เกรดเฉลี่ย และวิชาต่าง ๆ จากการกำหนดรูปแบบข้อมูลดังกล่าวจึงมีความเหมาะสม ต่อการนำเข้าโมเดลเพื่อทำนายแผนการเรียนมีจำนวนทั้งหมด 32 Attributes

ตารางที่ 3.4 แสดงตัวอย่างข้อมูลและ Code ของ Attributes แบบ Clustering

	O (1)	SEEM (2)	CEM (3)	B (4)	EM (5)	F (6)	N	NE	E	S	W	C
2	0	0	0	1	0	0	0	0	1	0	0	0
3	0	1	0	0	0	0	0	0	1	0	0	0
4	0	0	0	1	0	0	0	0	1	0	0	0
5	0	0	0	1	0	0	0	0	1	0	0	0
6	0	0	0	1	0	0	0	0	1	0	0	0
7	0	1	0	0	0	0	0	0	1	0	0	0
8	0	0	1	0	0	0	0	0	1	0	0	0
9	0	0	1	0	0	0	0	0	1	0	0	0
10	0	0	1	0	0	0	0	0	1	0	0	0
11	1	0	0	0	0	0	0	0	1	0	0	0
12	0	1	0	0	0	0	0	0	1	0	0	0
13	0	0	1	0	0	0	0	0	1	0	0	0
14	0	0	1	0	0	0	0	0	1	0	0	0
15	0	0	0	1	0	0	0	0	1	0	0	0
16	0	0	0	1	0	0	0	0	1	0	0	0
17	0	0	0	1	0	0	0	0	1	0	0	0
18	0	0	0	1	0	0	0	0	1	0	0	0
19	0	0	0	1	0	0	0	0	1	0	0	0
20	0	0	0	0	0	1	0	0	1	0	0	0

จากตารางที่ 3.4 Attributes แบบ Clustering แปลงค่าให้อยู่ในรูป 0 และ 1 ได้แก่ อาชีพผู้ปกครอง ภูมิลำเนา สำหรับเตรียมข้อมูล โดยมี Label แบ่งเป็น 2 ค่า คือ 1 = ใช่ และ0 = ไม่ใช่

ตารางที่ 3.5 แสดงตัวอย่างรูปแบบข้อมูลและ Code ข้อมูล

	BA	BB	BC	BD	BE	BF	BG	BH	BI	BJ	BK	BL	BM	BN
1	maths31	pe31	thai32	math32	sci32	soc32	his32	heal32	art32	we32	eng32	pe32	gpax3	gpax
2	1	4	2	1.5	1.5	2.5	2.5	4	3.5	4	2	4	2.98	2.65
3	1.5	4	3.5	2	3	3.5	4	4	3.5	4	3.5	4	3.17	2.78
4	2.5	4	3	3	3	2.5	3.5	4	4	4	2	4	3.28	2.67
5	1.5	4	2.5	2	1	1.5	2.5	4	2.5	3.5	2	4	2.03	2.01
6	2	4	3	2.5	2.5	2.5	2.5	4	3.5	4	2.5	4	3.2	2.86
7	3.5	4	3.5	4	4	4	4	4	4	4	4	4	3.85	3.65
8	3	4	3.5	3.5	3	3	3	4	2.5	4	3.5	4	3.11	3.6
9	1.5	4	2	1.5	1	2.5	3	4	3	4	2.5	4	2.91	2.13
10	2	4	3	4	3	3	3.5	4	3.5	4	4	4	3.56	3.38
11	3.5	4	3.5	3.5	3.5	3	3.5	4	2	4	4	4	3.65	3.52
12	4	4	4	4	4	4	4	4	1.5	4	4	4	3.89	3.87
13	3	4	2.5	1.5	2	1.5	3	4	3	4	1.5	4	2.93	2.96
14	2	4	2.5	2	2.5	3	3.5	4	2	4	1.5	4	3.21	2.85
15	3	4	3	3.5	3	3.5	2.5	4	3.5	4	3.5	4	3.57	3.39
16	3.5	4	3	2.5	3.5	3	2.5	4	3.5	4	3	4	3.39	3.35
17	4	4	3.5	4	4	4	3.5	4	4	4	4	4	3.9	3.78
18	4	4	3.5	4	4	4	4	4	4	3.5	4	4	3.89	3.94
19	4	4	3	4	3.5	3.5	4	4	4	4	4	4	3.73	3.19
20	2.5	4	2.5	3.5	3	2.5	2.5	4	2.5	4	3.5	4	3.2	3.03
21	2	4	3.5	3.5	3	3	3.5	4	2.5	4	4	4	3.19	3.4
22	2.5	4	4	3.5	3	4	4	4	4	4	3	4	3.53	3.58
23	2.5	4	3	3.5	3	3.5	3	4	3	4	3.5	4	3.48	3.65
24	2	4	3	2	3	3	3.5	4	1	4	3	4	2.8	2.8
25	3	4	4	4	4	4	4	4	2.5	4	4	4	3.84	3.92

จากตารางที่ 3.5 เมื่อดำเนินการรวบรวมข้อมูลและแปลงรูป เพื่อให้มีความเหมาะสมต่อการนำเข้าโมเดลวิเคราะห์แผนการเรียนด้วย วิธี Decision Tree, Naive Bayes, Artificial Neural network และ Support Vector Machine ได้ข้อมูลทั้งหมด 31 Attributes สำหรับการวิเคราะห์ความพันธ์ระหว่าง โดยมี Label คือ แผนวิทยาศาสตร์และคณิตศาสตร์ SM = 1 และแผนศิลป์ Art = 2

3.3.2 การเลือกข้อมูลเพื่อใช้ในการสร้าง Machine Learning

ทำการเลือกข้อมูลเพื่อใช้ในการสร้าง Machine Learning โดยตามนิยามนักเรียนที่ประสบผลสำเร็จในการเรียน หมายถึง นักเรียนที่ศึกษาในสายที่ตนเลือกเรียนต่อแล้วมีผลการเรียนเฉลี่ยเมื่อสำเร็จการศึกษาอยู่ที่ระดับ 3.00 ขึ้นไป โดยการนำข้อมูลเกรดเฉลี่ยของผู้สำเร็การศึกษาชั้นมัธยมศึกษาปีที่ 4 - 6 ปีการศึกษา 2554 - 2556 มาทำการเลือกผู้มีผลการศึกษาเกรดเฉลี่ยระดับ 3.00 ขึ้นไปเพื่อนำมาใช้กับ Machine Learning ผลการคัดเลือกจาก 454 ได้จำนวนข้อมูล 258 ชุด ดังภาพที่ 3.1

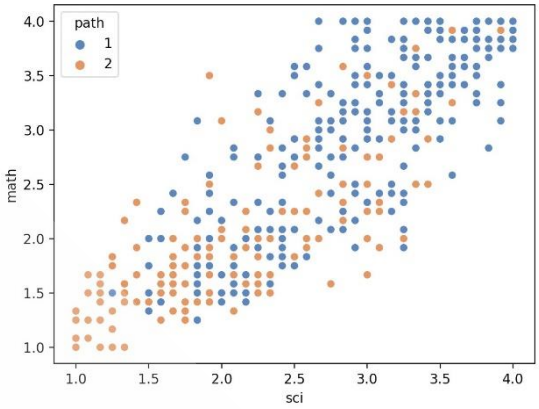
```
In [3]: df=pd.read_excel('highschoolchoosepath.xlsx')
        dcf=pd.read_excel('test_data.xlsx')
        df=df.loc[df['gpax']>=3.0]
        len(df)

Out[3]: 258
```

ภาพที่ 3.1 การเลือกข้อมูลเพื่อใช้ในการสร้าง Machine Learning

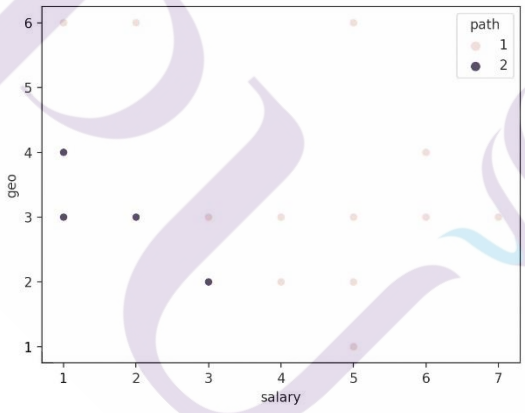
3.3.3 การวิเคราะห์ความสัมพันธ์ระหว่าง Attributes (Correlation Analysis)

นำข้อมูลมาวิเคราะห์ความสัมพันธ์ (Correlation Analysis) โดยการนำข้อมูลที่ผ่านกระบวนการมาวิเคราะห์ความสัมพันธ์ด้วยการสร้างกราฟ (Scatter Plot) เพื่อดูความน่าจะเป็นของข้อมูลการกระจายตัวโดยในงานวิจัยนี้ แบ่งกลุ่มแผนการเรียนต่อมัธยมศึกษาตอนปลาย ระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์หรือแผนศิลป์ การสร้าง Scatter Plot โดยในการวิเคราะห์ความสัมพันธ์นี้ใช้ Matplotlib [35] เป็น library การแสดงผลข้อมูล (Data Visualization) สร้างกราฟจาก Array กรอกค่าแสดงความสัมพันธ์ของตัวแปร ซึ่งสามารถสร้างกราฟในแบบ 2 มิติในภาษา Python เพื่อเป็นการตรวจสอบทำให้เกิดความมั่นใจว่าข้อมูลที่ผ่านกระบวนการกลั่นกรองเป็นชุดข้อมูลที่สามารถจะนำไปใช้ในการสร้างโมเดลในการวิเคราะห์แนวโน้มการเลือกแผนการเรียนมัธยมศึกษาตอนปลาย และเป็นข้อมูลที่เหมาะสมในการวิเคราะห์ เพื่อเปรียบเทียบประสิทธิภาพแบบจำลองที่สร้างขึ้น



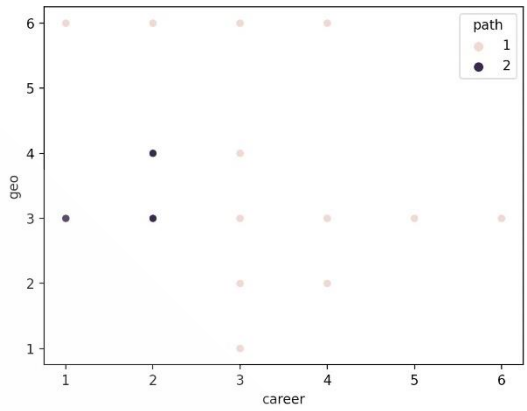
ภาพที่ 3.2 การวิเคราะห์ความสัมพันธ์ระหว่าง Attributes ด้วยการสร้าง Scatter Plot

จากภาพที่ 3.2 การวิเคราะห์ความสัมพันธ์ของชุดข้อมูลนี้จะเห็นการกระจายในลักษณะ 45 องศา ไปในทิศทางเดียวกัน แสดงว่าชุดข้อมูลนี้สามารถจะนำไปใช้ในการสร้างโมเดลในการวิเคราะห์แนวโน้มการเลือกแผนการเรียนมัธยมศึกษาตอนปลาย



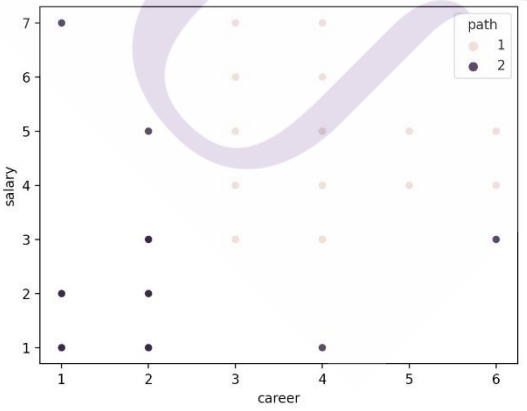
ภาพที่ 3.3 กราฟแสดงความสัมพันธ์ข้อมูลกลุ่มภูมิลำเนา

จากภาพที่ 3.3 แสดงลักษณะความสัมพันธ์ของชุดข้อมูล geo คือ ชุดข้อมูลภูมิลำเนา ประกอบด้วย ภาคเหนือ ภาคตะวันออกเฉียงเหนือ ภาคตะวันออก ภาคใต้ ภาคตะวันตก และภาคกลาง เป็นชุดข้อมูลที่ไม่ได้มีความสัมพันธ์กับข้อมูลรายได้ของผู้ปกครอง และของทั้งสองชุดนี้ไม่ได้มีความสัมพันธ์ที่จะส่งผลเพียงพอต่อการสร้างโมเดลวิเคราะห์แผนการเรียนด้วย วิธี Decision Tree, Naive Bayes, Artificial Neural network และ Support Vector Machine จะเห็นได้จากการกระจายตัวที่ไม่ได้เป็นไปในทิศทางเดียวกัน



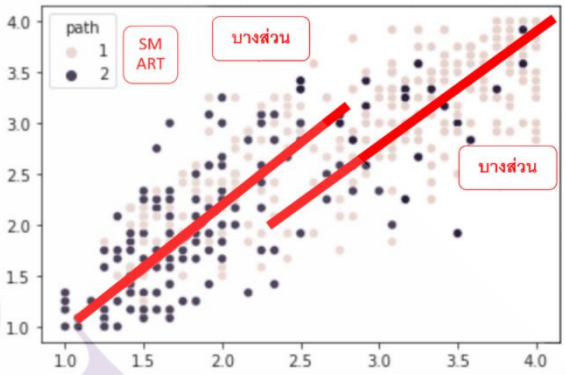
ภาพที่ 3.4 กราฟแสดงความสัมพันธ์ข้อมูลกลุ่มอาชีพของผู้ปกครอง

จากภาพที่ 3.4 จะแสดงลักษณะความสัมพันธ์ของชุดข้อมูล Career คือ ชุดข้อมูลอาชีพของผู้ปกครองประกอบด้วย ข้าราชการ/เจ้าหน้าที่ของรัฐ พนักงานรัฐวิสาหกิจ พนักงานบริษัทเอกชน ธุรกิจส่วนตัว/ค้าขาย รับจ้าง/ลูกจ้าง และเกษตรกรรม/ปศุสัตว์/ประมง เป็นชุดข้อมูลที่ไม่ได้มีความสัมพันธ์กับข้อมูลภูมิลำเนา และของทั้งสองชุดนี้ไม่ได้มีความสัมพันธ์ที่จะส่งผลเพียงพอต่อการสร้างโมเดลวิเคราะห์แผนการเรียนด้วย วิธี Decision Tree, Naive Bayes, Artificial Neural network และ Support Vector Machine จะเห็นได้จากการกระจายตัวที่ไม่ได้เป็นไปในทิศทางเดียวกัน



ภาพที่ 3.5 กราฟแสดงความสัมพันธ์ข้อมูลกลุ่มรายได้ของผู้ปกครอง

จากภาพที่ 3.5 จะแสดงลักษณะความสัมพันธ์ของชุดข้อมูล salary คือ รายได้ของผู้ปกครอง ประกอบด้วย รายได้ผู้ปกครอง น้อยกว่า 35,000 ช่วง50,001 - 65,000 ช่วง 65,001 - 80,000 ช่วง 80,001 - 95,000 ช่วง95,001 - 110,000 และ110,001 ขึ้นไป และของทั้งสองชุดนี้ไม่ได้มีความสัมพันธ์ที่จะส่งผลเพียงพอต่อการสร้างโมเดลวิเคราะห์แผนการเรียนด้วย วิธี Decision Tree, Naive Bayes, Artificial Neural network และ Support Vector Machine จะเห็นได้จากการกระจายตัวที่ไม่ได้เป็นไปในทิศทางเดียวกัน



ภาพที่ 3.6 การวิเคราะห์ความสัมพันธ์ ด้วยการสร้าง Scatter Plot

จากภาพที่ 3.6 เป็นการวิเคราะห์ความสัมพันธ์ ด้วยการสร้าง Scatter Plot โดยใช้ Matplotlib แสดงให้เห็นถึงการกระจายตัวและการเกาะกลุ่มของข้อมูลที่เราได้ขจัดชุดข้อมูลที่ไม่ได้มีความสัมพันธ์กับกลุ่มข้อมูลหลัก คือ กลุ่มข้อมูลอาชีพ รายได้ ภูมิภาค จำนวนทั้งหมด 19 Attributes ทำให้เหลือ 12 Attributes ที่ผู้วิจัยจะกำหนดให้เป็นข้อมูลชุดที่ 1 ใช้ในการวิจัย สร้างโมเดลในการทำนายด้วยการแบ่งข้อมูลออกเป็นสองกลุ่ม ตามแผนการศึกษาต่อมัธยมศึกษาตอนปลาย ระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์ (SM) แสดงด้วยจุดสีอ่อน และแผนศิลป์ (ART) โดยแสดงด้วยจุดสีเข้มและจุดสีอ่อน จากการวิเคราะห์แสดงให้เห็นถึงการเกาะกลุ่มกันของข้อมูลอย่างชัดเจนโดยกลุ่มแผนวิทยาศาสตร์และคณิตศาสตร์ (SM) เกาะกลุ่มอยู่ทางบนด้านขวา และแผนศิลป์ (ART) เกาะกลุ่มอยู่ทางด้านซ้ายล่าง ทั้งสองกลุ่มแยกออกจากกันและมีการกระจายตัวลักษณะทิศทาง 45 องศา จึงเป็นแสดงให้เห็นว่าชุดข้อมูลที่เราจะนำมาสร้างและทดสอบประสิทธิภาพของโมเดลมีแนวโน้มที่จะสามารถนำมาทำนายแผนการเรียนต่อมัธยมศึกษาตอนปลายได้จึงกำหนดให้เป็นข้อมูลชุดที่ 1 จำนวน 12 Attributes

3.3.4 การวิเคราะห์การลดขนาดข้อมูล (Principal Component Analysis : PCA)

หลังจากวิเคราะห์ความสัมพันธ์ (Correlation Analysis) โดยการนำข้อมูลที่ผ่านกระบวนการมาวิเคราะห์ความสัมพันธ์ด้วยการสร้างกราฟ (Scatter Plot) ผู้วิจัยทำการวิเคราะห์โดยวิธีการลดขนาดข้อมูล (Principal Component Analysis : PCA) เพื่อจะกำหนดให้เป็นข้อมูลชุดที่ 2 โดยทำการวิเคราะห์องค์ประกอบหลัก PCA เป็นการลดความแปรปรวนและเพิ่มประสิทธิภาพในการนำข้อมูลไปวิเคราะห์ เพื่อให้มีความเหมาะสม ต่อการนำเข้าโมเดลมาวิเคราะห์แผนการเรียนด้วย วิธี Decision Tree, Naive Bayes, Artificial Neural network และ Support Vector Machine โดย PCA เป็นขั้นตอนทางสถิติที่แปลงมิติสมการทางคณิตศาสตร์จากชุดข้อมูลที่เราต้องการวิเคราะห์ให้เป็นชุดข้อมูลใหม่ที่เรียกว่าองค์ประกอบหลัก อันเป็นผลมาจากการเปลี่ยนแปลงองค์ประกอบรวมทำให้เหลือ Attributes ที่มีความสัมพันธ์กัน และทำให้จำนวนน้อยลงด้วยการดูจากความแปรปรวนที่มากที่สุดขององค์ประกอบ ช่วยให้เราสามารถลดมิติข้อมูลในขณะที่ยังคงรักษาข้อมูลส่วนใหญ่ที่จะมีผลกับการนำไปวิเคราะห์ โดยหลังจากทำ PCA แล้ว ได้ลดรูปเกิดเป็นองค์ประกอบใหม่จำนวน 4 Attributes โดยมีขั้นตอนดังนี้

1) เป็นการเริ่มต้นการเรียกใช้งาน PCA ที่อยู่ในตัว Decomposition และทำการสเกลข้อมูลที่อยู่ในฐานข้อมูลที่เราจะให้ในการวิเคราะห์การลดขนาดข้อมูล ให้มีค่าน้ำหนักเท่ากันเช่น คะแนนรายวิชากับเพศ หรือรายได้มีหน่วยที่ต่างกัน จึงต้องตั้งให้อยู่ในสเกลเดียวกันก่อนการวิเคราะห์ ด้วยคำสั่ง Stand Scaler ในรูปของ z – score ดังภาพที่ 3.7

PCA

```
1 from sklearn.decomposition import PCA
2 from sklearn.preprocessing import StandardScaler # z-score
```

ภาพที่ 3.7 การเริ่มต้น PCA และการทำสเกลข้อมูล

2) ผลของ PCA วิเคราะห์ความแปรปรวนจาก PCA สร้างตัว Object ขึ้นมาโดยบอกจำนวน Components ในที่นี้คือ x_t หมายถึง ข้อมูลที่ Transform มาแล้ว หลังจากนั้น PCA ก็จะส่งค่ากลับมาใน Scikit - learn เรียกว่า Explained variance หรือ Eigenvector ในการสร้าง Components ใหม่ครั้งนี้ได้ทำให้ข้อมูลจำนวน 12 Attributes สร้างองค์ประกอบใหม่เป็น 4 Components เพื่อใช้เป็นข้อมูลชุดที่ 2 ดังภาพที่ 3.8

```
pca = PCA(n_components=X_t.shape[1])
# pca = PCA(n_components)

pca.fit_transform(X_t)
print(f'explained_variance (n_components={pca.n_components}) = {pca.explained_variance_}') # Eigenv
print(f'explained_variance_ratio (n_components={pca.n_components}) = {pca.explained_variance_ratio_
print(f'sum explained_variance_ratio = {np.sum(pca.explained_variance_ratio_)}')
```

ภาพที่ 3.8 ผลของ PCA จากค่า Variance

3.3.5 การวิเคราะห์ตัวเลือกคุณลักษณะตามลำดับ (Feature Selector : Backward Selection)

ผู้วิจัยกำหนดข้อมูลชุดที่ 3 โดยวิธี Backward Selection เป็นการคัดเลือก Attributes โดยการนำ Attributes ทุกตัวเข้ามาในสมการและดำเนินการพิจารณา Attributes ที่มีค่าสัมประสิทธิ์สหสัมพันธ์บางส่วน (Partial Correlation) กับเกณฑ์ โดยควบคุมอิทธิพลของ Attributes อื่น ๆ ซึ่งมีค่าต่ำที่สุดออกจากสมการ แล้วทดสอบว่าค่าที่ลดลงอย่างมีนัยสำคัญทางสถิติหรือไม่ ถ้าพบว่าลดลงอย่างไม่มีนัยสำคัญทางสถิติแสดงว่า Attributes ดังกล่าวไม่ได้ทำให้การทำนายเพิ่มขึ้น แสดงว่าสามารถขจัดออกจากสมการได้เราจึงดำเนินการขจัด Attributes ที่มีความสำคัญน้อยรองลงมาออกไป จะทำการขจัด Attributes ไป จนพบว่ามีผลทำให้ค่าที่ลดลงอย่างมีนัยสำคัญทางสถิติดังนั้น Attributes ที่มีความสำคัญต่อการทำนาย เมื่อขจัดตัวแปรดังกล่าวออกจากสมการจะทำให้อำนาจการทำนายลดลงจึงต้องคง Attributes ที่เหลือเอาไว้ในการสร้างการทำนายโมเดลวิเคราะห์แผนการเรียนด้วยวิธี Decision Tree, Naive Bayes, Artificial Neural network และ Support Vector Machine ที่ใช้ Backward Selection จะมีการเลือก Attributes ที่แตกต่างกันในแต่ละวิธีดังนี้ วิธี 1)Decision Tree วิธี Backward Selection คัดเลือก Attributes (math, sci, soc, his, pe) 2)Artificial Neural Network วิธี Backward Selection คัดเลือก Attributes (math, eng, soc, heal, his) 3)Naive Bayes วิธี Backward Selection คัดเลือก Attributes (sex, math, soc, his, gpax) 4)Support Vector Machine วิธี Backward Selection คัดเลือก Attributes (math, sci, eng, pe, gpax)

```
clf = tree.DecisionTreeClassifier(random_state=rs,criterion='gini')
#sfs = SequentialFeatureSelector(clf,n_features_to_select=11,direction='backward')
sfs = SequentialFeatureSelector(clf,direction='backward')
sfs.fit(X_train, y_train)
```

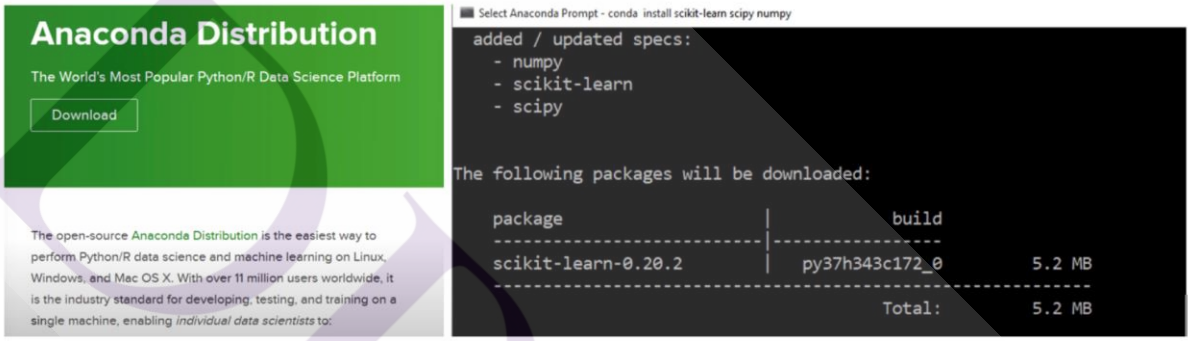
ภาพที่ 3.9 การใช้คำสั่ง Feature Selector: Backward Selection

3.4 การสร้างแบบจำลอง (Model) ทดสอบและเปรียบเทียบประสิทธิภาพ

การสร้างแบบจำลองออกแบบโดยใช้ Scikit - learn บนภาษา Python มีการทำงานแบบการแปลชุดคำสั่งทีละบรรทัด เลือกใช้ Machine Learning ในแบบ Supervised Learning ประกอบด้วย 4 วิธีคือ 1) Decision Tree 2) Artificial Neural Network 3) Naive Bayes 4) Support Vector Machine ดังนี้

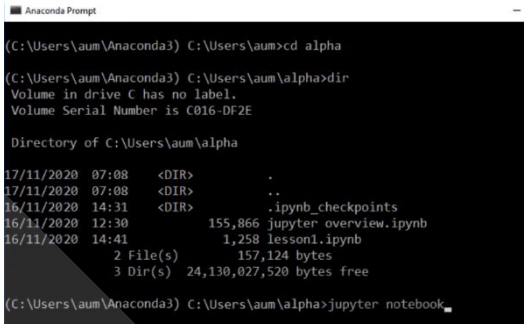
3.4.1 เตรียม Machine Learning

ผู้วิจัยเริ่มต้นด้วยการติดตั้ง Python จาก Anaconda Distribution ซึ่งภายใน Anaconda นี้จะมีการ Package Scikit-learn ให้ แล้วก็สามารถติดตั้งได้ทันทีดังภาพที่ 3.10



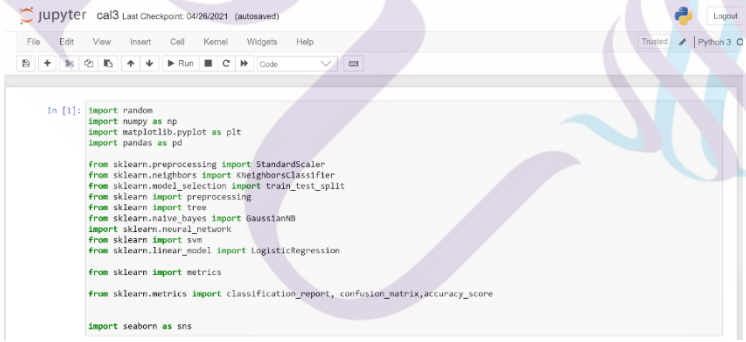
ภาพที่ 3.10 การ Download และการติดตั้ง Anaconda

ลำดับต่อมาผู้วิจัยเรียกใช้ Integrated Development Environment ที่ชื่อว่า Jupyter notebook เริ่มจากการเรียก Anaconda Prompt แล้วค้นหา Directory ที่เก็บ notebook และเริ่มต้นโดยการพิมพ์ Jupyter notebook ก็จะปรากฏขึ้นมา Jupyter นี้ใช้สำหรับป้อนคำสั่งเพื่อประมวลผลคำสั่งนั้น ๆ พร้อมแสดงผลลัพธ์ต่อท้ายชุดคำสั่ง สามารถเก็บ code ตัว markdown และ Raw ที่ช่วยให้แสดงข้อมูลในแบบต่าง ๆ ได้ เช่น แสดงเป็น Heading ตัว Jupyter notebook สามารถใช้งานได้ง่ายทั้งผ่านการพิมพ์และการใช้งานผ่านตัว Tool bar หรือ Menu ก็ได้และสามารถเข้าถึงผ่าน Command palette เช่นการใช้ code โดยการกด Ctrl Shift P ก็จะเปิด Command ขึ้นมา ดังตัวอย่างภาพที่ 3.11



ภาพที่ 3.11 Integrated Development Environment

เริ่มต้นการใช้ Jupyter notebook ในการทำ Machine Learning ด้วยการ Import Library ประกอบด้วย Numpy ใช้ฟังก์ชั่นทางคณิตศาสตร์และการคำนวณ Matplotlib ใช้การแสดงผลข้อมูล (Data Visualization) ด้วยการสร้างกราฟ Pandas ใช้ในการจัดการข้อมูล (Data Manipulation) ช่วยจัดการข้อมูลที่จะต้อง import และแสดงข้อมูลสามารถอ่านไฟล์ได้หลากหลาย และเริ่มทำมาโหลดวิธี Decision Tree, Naive Bayes , Artificial Neural network และ Support Vector Machine มารอเพื่อเรียกใช้ในการสร้างโมเดลต่อไปดังภาพที่ 3.12



ภาพที่ 3.12 การใช้ Jupyter notebook ในการทำ Machine Learning

3.4.2 การสร้างแบบจำลอง (Model)

1) Decision Tree งานวิจัยนี้ใช้ การสร้างแบบจำลองออกแบบ โดยใช้ Scikit-learn บนภาษา Python ใน Scikit-learn นั้นใช้ Algorithm สำหรับเลือก Root Node ที่ชื่อว่า CART (Classification and Regression Trees) โดยมีลักษณะเด่นที่รองรับตัวแปรเป้าหมายที่เป็นตัวเลข (Regression) แบบ Random Generator และไม่ Compute rule sets กฎ CART สร้าง Binary Trees

โดยใช้คุณลักษณะและขีดจำกัด ข้อมูลได้รับมากที่สุดในแต่ละโหนด และ CART มีความคล้ายกับ C4.5 โดยเริ่มจากการแบ่งข้อมูลชุดเรียนรู้ (Training set) ร้อยละ 80 และข้อมูลชุดทดสอบ (Testing set) ร้อยละ 20 และ ใช้ Gini ในการจำแนกในการหา Root Node และใช้ค่าในการตรวจสอบโมเดล คือ Recall, Precision และ F1-score ดังภาพที่ 3.13

```
X_train, X_test, y_train, y_test = train_test_split(x_data, y_data, test_size=0.2,random_state=rs)
clf = tree.DecisionTreeClassifier(random_state=10,criterion='gini')
clf = clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
print('accuracy: {:.4f}'.format(accuracy_score(y_test,y_pred)))
```

ภาพที่ 3.13 การสร้างโมเดล Decision Tree

2) Artificial Neural Network ในงานวิจัยนี้ตั้งค่าโดยเริ่มจากการแบ่งข้อมูลชุดเรียนรู้ (Training Set) ร้อยละ 80 และข้อมูลชุดทดสอบ (Testing set) ร้อยละ 20 และกำหนด 4 Hidden layer และกำหนด Hidden layer sizes = 50 ในการทำโมเดลนี้มีการแปรผลค่า และใช้ค่าในการตรวจสอบโมเดลคือ Recall, Precision และ F1 - score ดังภาพที่ 3.14

```
X_train, X_test, y_train, y_test = train_test_split(x_data, y_data, test_size=0.2,random_state=10)
model = sklearn.neural_network.MLPClassifier(hidden_layer_sizes=(50,50,50,50), activation='relu', solver='adam',
                                             alpha=0.0001, batch_size='auto', learning_rate='constant', learning_ra
                                             max_iter=100000, shuffle=True, random_state=True, tol=0.0001, verbose=
                                             nesterovs_momentum=True, early_stopping=False, validation_fraction=0.1
                                             n_iter_no_change=10)

model.fit(X_train, y_train)
y_predANN = model.predict(X_test)
print(confusion_matrix(y_test, y_predANN))
print(classification_report(y_test, y_predANN))
print('accuracy: {:.4f}'.format(accuracy_score(y_test,y_predANN)))
```

ภาพที่ 3.14 การสร้างโมเดล Artificial Neural Network

3) Naive Bayes งานวิจัยนี้ตั้งค่า โดยเริ่มจากการแบ่งข้อมูลชุดเรียนรู้ (Training set) ร้อยละ 80 และข้อมูลชุดทดสอบ (Testing set) ร้อยละ 20 ใน Scikit-learn นั้นใช้ Naive Bayes Algorithms ชื่อว่า Gaussian Naive Bayes มีคุณลักษณะที่ใช้ค่าต่อเนื่อง เป็นไปตามการแจกแจงแบบ Gaussian ตามกฎของเบย์ การแจกแจงแบบปกติ ในการทำโมเดลนี้มีการแปรผลค่า และใช้ค่าในการตรวจสอบโมเดลคือ Recall, Precision และ F1-score ดังภาพที่ 3.15

```
X_train, X_test, y_train, y_test = train_test_split(x_data, y_data, test_size=0.2,random_state=10)
gnb = GaussianNB()
y_predgnb = gnb.fit(X_train, y_train).predict(X_test)
print(confusion_matrix(y_test, y_predgnb))
print(classification_report(y_test, y_predgnb))
print('accuracy: {:.4f}'.format(accuracy_score(y_test,y_predgnb)))
```

ภาพที่ 3.15 การสร้างโมเดล Naive Bayes

4) Support Vector Machine ในงานวิจัยนี้ตั้งค่า โดยเริ่มจากการแบ่งข้อมูลชุดเรียนรู้ (Training set) ร้อยละ 80 และข้อมูลชุดทดสอบ (Testing set) ร้อยละ 20 ในการทำโมเดลนี้มีการแปรผลค่าและใช้ค่าในการตรวจสอบโมเดลคือ Recall, Precision และ F1-score ดังภาพที่ 3.16

```
X_train, X_test, y_train, y_test = train_test_split(x_data, y_data, test_size=0.2,random_state=10)
csvm=svm.SVC()
csvm.fit(X_train,y_train)
y_predsvm = csvm.predict(X_test)
print(confusion_matrix(y_test, y_predsvm))
print(classification_report(y_test, y_predsvm))
print('accuracy: {:.4f}'.format(accuracy_score(y_test,y_predsvm)))
```

ภาพที่ 3.16 การสร้างโมเดล Support Vector

3.5 การแปรผลและการประเมินผล (Interpretation Evaluation)

นำเสนอผลการทดสอบประสิทธิภาพแบบจำลองทั้ง 4 วิธี 1)Decision Tree 2)Artificial Neural Network 3)Naive Bayes 4)Support Vector Machine แสดงค่าต่อไปนี้

1) ค่าความแม่นยำ (Precision) ของการเลือกระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์ ซึ่งแทนด้วยเลข 1 หรือแผนศิลป์แทนด้วยเลข 2 แสดงค่าจากตารางด้วยค่าร้อยละตามลำดับ แทนค่าในสมการ Precision = TP / (TP + FP) ค่า TP หมายถึง ข้อมูลที่ทำนายแล้วถูกต้องเมื่อเทียบกับคำตอบ และ FP หมายถึงข้อมูลที่อยู่ในคำตอบแต่ไม่มีในการทำนาย

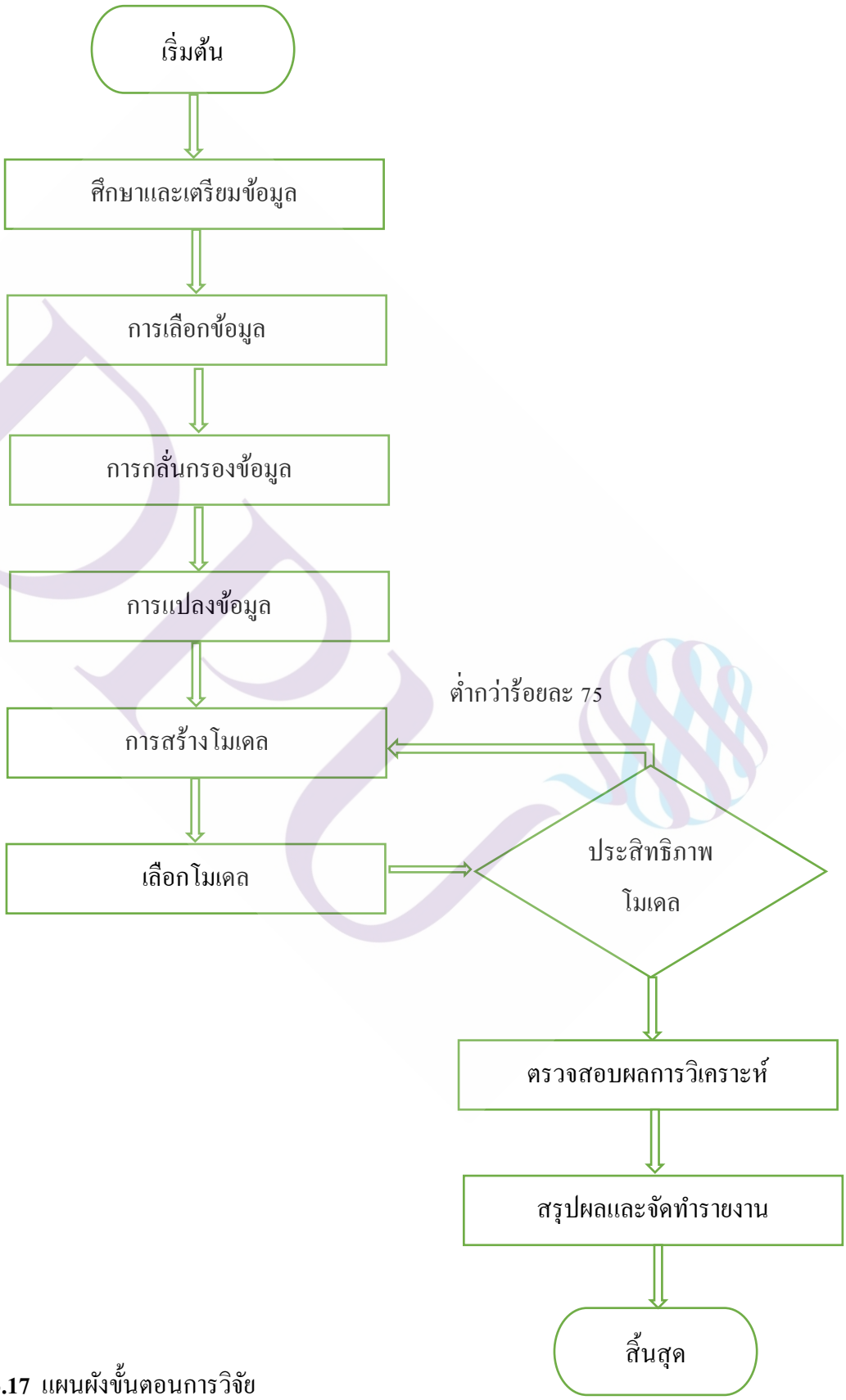
2) ค่าโอกาสการเรียกซ้ำ (Recall) แผนการเรียนที่ 1 หรือ 2 แสดงค่าจากตารางด้วยค่าร้อยละตามลำดับในสมการ Recall = TP / (TP + FN) ค่า FN หมายถึง ข้อมูลที่ทำนายแล้วไม่ถูกต้องเมื่อเทียบกับคำตอบ

3) F1- Score เป็นการตรวจสอบความแม่นยำของแบบจำลอง คือ ค่าเฉลี่ยแบบ Harmonic Mean ระหว่าง Precision และ Recall เป็น Single Metric ที่วัดความสามารถของโมเดล แทนค่าในสมการ F1 = 2 * ((Precision * Recall) / (Precision + Recall))

4) จำนวนครั้งที่จะเกิดการเลือก (Support) แผนการเรียนที่ 1 และ 2 รวมกัน

5) ค่าร้อยละความถูกต้อง (Accuracy) แสดงค่าที่โมเดลทำนายถูกทั้งหมดเป็นค่าร้อยละ

ขั้นตอนการสร้างแบบจำลอง



ภาพที่ 3.17 แผนผังขั้นตอนการวิจัย

บทที่ 4
ผลดำเนินการวิจัย

การรายงานผลการวิจัย จะเป็นการแสดงผลการทดลองของแบบจำลองที่ถูกพัฒนาขึ้นเพื่อเป็นเครื่องมือช่วยตัดสินใจเลือกแผนการเรียนต่อมัธยมศึกษาตอนปลายระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์หรือแผนศิลป์ โดยเลือกใช้การเรียนรู้ของเครื่องในแบบ Supervised Learning ประกอบด้วย

1) Decision Tree

2) Artificial Neural Network

3) Naive Bayes

4) Support Vector Machine

การสร้างและหาประสิทธิภาพแบบจำลองทั้ง 4 แบบ จะแสดงค่าการเปรียบเทียบในแต่ละแบบว่ามีประสิทธิภาพแตกต่างกันอย่างไร และในท้ายบทจะมีค่าประสิทธิภาพแบบจำลองทั้ง 4 แบบ ทดลองด้วยข้อมูล 3 ชุด เริ่มจากข้อมูลชุดที่ 1 เป็นชุดข้อมูลที่ผ่านกระบวนการ Correlation Analysis มีข้อมูลจำนวน 12 Attributes ข้อมูลชุดที่ 2 เป็นชุดข้อมูลที่ใช้วิธีการลดขนาดข้อมูล (Principal Component Analysis : PCA) มีข้อมูลจำนวน 4 Attributes ข้อมูลชุดที่ 3 ใช้วิธีการวิเคราะห์ตัวเลือกคุณลักษณะตามลำดับ Feature Selector (Sequential Feature Selector : Backward Selection) จะมีการเลือก Attributes ที่แตกต่างกันในแต่ละ Algorithms ที่ใช้

ผู้วิจัยเสนอผลการวิเคราะห์ข้อมูลแบ่งเป็น 3 ตอน ดังนี้

ตอนที่ 1 ผลการวิเคราะห์ประสิทธิภาพโมเดลของข้อมูลชุดที่ 1 ชุดที่ 2 และชุดที่ 3

ตอนที่ 2 การเปรียบเทียบผลการวิเคราะห์ประสิทธิภาพโมเดลด้วยข้อมูลชุดที่ 1 ชุดที่ 2 และชุดที่ 3

ตอนที่ 3 สรุปผลการวิเคราะห์

จำนวนข้อมูลเพื่อใช้ในการสร้าง Machine Learning

นำข้อมูลเกรดเฉลี่ยของผู้สำเร็จการศึกษาชั้นมัธยมศึกษาปีที่ 4 - 6 ปีการศึกษา 2554 - 2556 มาทำการเลือกจากผู้มีผลการเรียนเกรดเฉลี่ยระดับ 3.00 ขึ้นไป เพื่อใช้ในการสร้าง Machine Learning ผลการคัดเลือกจาก 454 ได้จำนวนข้อมูล 258 ชุด

- ตอนที่ 1 ผลการวิเคราะห์ประสิทธิภาพโมเดลของข้อมูลชุดที่ 1 ชุดที่ 2 และชุดที่ 3

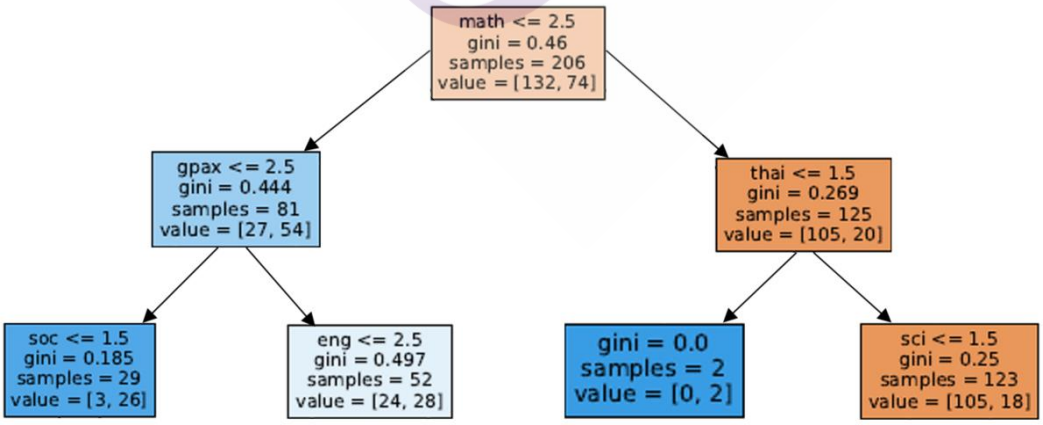
1.1 การทดสอบแบบจำลองชุดที่ 1 ด้วยเทคนิค Decision Tree

จากแบบจำลองด้วยเทคนิค Decision Tree เริ่มต้นด้วยแบ่งข้อมูลชุดเรียนรู้ (Training Set) ร้อยละ 80 และข้อมูลชุดทดสอบ (Testing set) ร้อยละ 20 จากนั้นกำหนดค่าการเลือกระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์ (SM) ซึ่งแทนด้วยเลข 1 หรือแผนศิลป์ (ART) แทนด้วยเลข 2

ตารางที่ 4.1 ผลการทดสอบแบบจำลองชุดที่ 1 ด้วยเทคนิค Decision Tree

Decision Tree	Correlation Analysis			
	Precision	Recall	F1 - score	Support
SM 1	0.82	0.89	0.85	36
ART 2	0.69	0.56	0.62	16
Accuracy			0.79	52

จากตารางที่ 4.1 มีค่าความแม่นยำ (Precision) ของการเลือกระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์หรือแผนศิลป์ ร้อยละ 82 และ69 ตามลำดับ ค่าโอกาสการเรียกซ้ำ (Recall) แผนการเรียนที่ 1 และแผนการเรียนที่ 2 ที่ร้อยละ 89 และ56 ตามลำดับ ผลที่ได้จากแบบจำลองด้วยเทคนิค Decision Tree จะใช้การตรวจสอบ แบบ F1-score เพื่อตรวจสอบความแม่นยำของแบบจำลองอยู่ที่ ร้อยละ 85 และ62 จำนวนครั้งที่จะเกิดการเลือกแผนการเรียนที่ 1 และ2 (Support) รวมกันที่ 52 โดยค่าร้อยละความถูกต้อง (Accuracy) ของ Decision Tree มีความถูกต้องที่ร้อยละ 79



ภาพที่ 4.1 แผนผังต้นไม้จากการทำนายข้อมูลชุดที่ 1 ด้วยเทคนิค Decision Tree

จากภาพที่ 4.1 อธิบายลักษณะการทำนายผลจากข้อมูลชุดที่ 1 Correlation Analysis ด้วยเทคนิค Decision Tree จากข้อมูล 258 ชุด Machine Learning จะสามารถหา Root node ได้โดยค่า Gini ที่มีค่าสูงสุดในการทำนายนี้ เริ่มต้นจากคำถามแรกจากค่าเกรดเฉลี่ยของวิชาคณิตศาสตร์เป็นตัวเริ่มต้นแสดงให้เห็นว่าใน Attributes ทั้งหมด วิชาคณิตศาสตร์เป็นตัวจำแนกได้ดีที่สุด โดยทำนายได้ 206 จาก 258 คน และคำถามคือ ค่าคะแนนเฉลี่ยของวิชาคณิตศาสตร์มากหรือน้อยกว่า 2.5 หากน้อยกว่าให้เลือกถามคำถามเรื่องคะแนนเฉลี่ยมัธยมศึกษาตอนต้นด้านซ้ายมือ แต่ถ้ามากกว่า 2.5 ให้ถามคำถามที่ว่า คะแนนวิชาภาษาไทยมากกว่าหรือน้อยกว่า 1.5 เพื่อเลือกคำถามถัดไปจนจบสายของแผนผังต้นไม้จากการทำนายด้วยเทคนิค Decision Tree

1.2 การทดสอบแบบจำลองข้อมูลชุดที่ 1 ด้วยเทคนิค Artificial Neural Network

เมื่อทดลองแบบจำลองด้วยเทคนิค Artificial Neural Network โดยแบ่งข้อมูลชุดเรียนรู้ (Training set) ร้อยละ 80 และข้อมูลชุดทดสอบ (Testing set) ร้อยละ 20 ค่าความแม่นยำ (Precision) ของการเลือกระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์ (SM) แทนด้วยเลข 1 หรือแผนศิลป์ (ART) แทนด้วยเลข 2

ตารางที่ 4.2 ผลการทดสอบแบบจำลองข้อมูลชุดที่ 1 ด้วยเทคนิค Artificial Neural Network

ANN	**Correlation Analysis**			
	Precision	Recall	F1 - score	Support
SM 1	0.87	0.89	0.88	37
ART 2	0.71	0.67	0.69	15
Accuracy			0.83	52

จากตารางที่ 4.2 ค่าความแม่นยำ (Precision) ของการเลือกระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์หรือแผนศิลป์ มีค่าร้อยละ 87 และ71 ตามลำดับ ค่าโอกาสการเรียกซ้ำ (Recall) แผนการเรียนที่ 1 หรือ 2 ที่ร้อยละ 89 และ67 ตามลำดับ ผลที่ได้จากแบบจำลองด้วยเทคนิค Artificial Neural Network จะใช้การตรวจสอบ แบบ F1-score เพื่อตรวจสอบความแม่นยำของแบบจำลองอยู่ที่ ร้อยละ 88 และ69 จำนวนครั้งที่จะเกิดการเลือกแผนการเรียนที่ 1 และ2 (Support) รวมกันที่ 52 โดยค่าร้อยละความถูกต้อง (Accuracy) ของ Artificial Neural Network มีความถูกต้องที่ร้อยละ 83

1.3 การทดสอบแบบจำลองข้อมูลชุดที่ 1 ด้วยเทคนิค Naive Bayes

จากการทดลองใช้แบบจำลองด้วยเทคนิค Naive Bayes เมื่อแบ่งข้อมูลชุดเรียนรู้ (Training set) ร้อยละ 80 และข้อมูลชุดทดสอบ (Testing set) ร้อยละ 20 ค่าความแม่นยำ (Precision) ของการเลือกระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์ (SM) แทนด้วยเลข 1 หรือแผนศิลป์ (ART) แทนด้วยเลข 2

ตารางที่ 4.3 ผลการทดสอบแบบจำลองข้อมูลชุดที่ 1 ด้วยเทคนิค Naive Bayes

Naive Bayes	**Correlation Analysis**			
	Precision	Recall	F1 - score	Support
SM 1	0.88	0.95	0.91	37
ART 2	0.83	0.67	0.74	15
Accuracy			0.87	52

จากตารางที่ 4.3 ค่าความแม่นยำ (Precision) ของการเลือกระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์หรือแผนศิลป์ มีค่าร้อยละ 88 และ83 ตามลำดับ ค่าโอกาสการเรียกซ้ำ (Recall) แผนการเรียนที่ 1 หรือ 2 ที่ร้อยละ 95 และ67 ตามลำดับ ผลที่ได้จากแบบจำลองด้วยเทคนิค Naive Bayes จะใช้การตรวจสอบ แบบ F1-score เพื่อตรวจสอบความแม่นยำของแบบจำลองอยู่ที่ ร้อยละ 91 และ74 จำนวนครั้งที่จะเกิดการเลือกแผนการเรียนที่ 1 และ2 (Support) รวมกันที่ 52 โดยค่าร้อยละความถูกต้อง (Accuracy) ของ Naive Bayes มีความถูกต้องที่ร้อยละ 87

1.4 การทดสอบแบบจำลองข้อมูลชุดที่ 1 ด้วยเทคนิค Support Vector Machine

เมื่อใช้แบบจำลองด้วยเทคนิค Support Vector Machine โดยแบ่งข้อมูลชุดเรียนรู้ (Training Set) ร้อยละ 80 และข้อมูลชุดทดสอบ (Testing set) ร้อยละ 20 ค่าความแม่นยำ (Precision) ของการเลือกระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์ (SM) แทนด้วยเลข 1 หรือแผนศิลป์ (ART) แทนด้วยเลข 2

ตารางที่ 4.4 ผลการทดสอบแบบจำลองข้อมูลชุดที่ 1 ด้วยเทคนิค Support Vector Machine

Support Vector Machine	**Correlation Analysis**			
	Precision	Recall	F1 - score	Support
SM 1	0.86	0.94	0.90	34
ART 2	0.87	0.72	0.79	18
Accuracy			0.87	52

จากตารางที่ 4.4 ค่าความแม่นยำ (Precision) ของการเลือกระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์หรือแผนศิลป์ มีค่าร้อยละ 86 และ 87 ตามลำดับ ค่าโอกาสการเรียกซ้ำ (Recall) แผนการเรียนที่ 1 หรือ 2 ที่ร้อยละ 94 และ72 ตามลำดับ ผลที่ได้จากแบบจำลองด้วยเทคนิค Support Vector Machine จะใช้การตรวจสอบ แบบ F1-score เพื่อตรวจสอบความแม่นยำของแบบจำลองอยู่ที่ ร้อยละ 90 และ79 จำนวนครั้งที่จะเกิดการเลือกแผนการเรียนที่ 1 และ2 (Support) รวมกันที่ 52 โดยค่าร้อยละความถูกต้อง (Accuracy) ของ Support Vector Machine มีความถูกต้องที่ร้อยละ 87

1.5 การทดสอบแบบจำลองชุดที่ 2 ด้วยเทคนิค Decision Tree

จากแบบจำลองด้วยเทคนิค Decision Tree ใช้วิธีการลดขนาดข้อมูล (Principal Component Analysis : PCA) เริ่มต้นด้วยแบ่งข้อมูลชุดเรียนรู้ (Training Set) ร้อยละ 80 และข้อมูลชุดทดสอบ (Testing set) ร้อยละ 20 จากนั้นกำหนดค่าการเลือกระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์ (SM) ซึ่งแทนด้วยเลข 1 หรือแผนศิลป์ (ART) แทนด้วยเลข 2

ตารางที่ 4.5 ผลการทดสอบแบบจำลองชุดที่ 2 ด้วยเทคนิค Decision Tree

Decision Tree	**PCA**			
	Precision	Recall	F1 - score	Support
SM 1	0.80	0.92	0.86	36
ART 2	0.73	0.50	0.59	16
Accuracy			0.79	52

จากตารางที่ 4.5 มีค่าความแม่นยำ (Precision) ของการเลือกระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์หรือแผนศิลป์ร้อยละ 80 และ73 ตามลำดับ ค่าโอกาสการเรียกซ้ำ (Recall) แผนการเรียนที่ 1 และแผนการเรียนที่ 2 ที่ร้อยละ 92 และ50 ตามลำดับ ผลที่ได้จากแบบจำลองด้วยเทคนิค Decision Tree จะใช้การตรวจสอบ แบบ F1-score เพื่อตรวจสอบความแม่นยำของแบบจำลองอยู่ที่ร้อยละ 86 และ59 จำนวนครั้งที่จะเกิดการเลือกแผนการเรียนที่ 1 และ2 (Support) รวมกันที่ 52 โดยค่าร้อยละความถูกต้อง (Accuracy) ของ Decision Tree มีความถูกต้องที่ร้อยละ 79 สำหรับแผนผังของ Decision Tree ด้วยชุดข้อมูล PCA นี้จะแสดงเป็นลักษณะความเชื่อมโยงระหว่าง 4 Components ทำให้การอธิบายความสัมพันธ์ทำได้ยากเพราะองค์ประกอบถูกรวมกันไปแล้ว

1.6 การทดสอบแบบจำลองข้อมูลชุดที่ 2 ด้วยเทคนิค Artificial Neural Network

เมื่อทดลองแบบจำลองด้วยเทคนิค Artificial Neural Network ใช้วิธีการลดขนาดข้อมูล (Principal Component Analysis : PCA) โดยแบ่งข้อมูลชุดเรียนรู้ (Training set) ร้อยละ 80 และข้อมูลชุดทดสอบ (Testing set) ร้อยละ 20 ค่าความแม่นยำ (Precision) ของการเลือกระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์ (SM) แทนด้วยเลข 1 หรือแผนศิลป์ (ART) แทนด้วยเลข 2

ตารางที่ 4.6 ผลการทดสอบแบบจำลองข้อมูลชุดที่ 2 ด้วยเทคนิค Artificial Neural Network

ANN	**PCA**			
	Precision	Recall	F1 - score	Support
SM 1	0.82	1.00	0.90	37
ART 2	1.00	0.47	0.64	15
Accuracy			0.85	52

จากตารางที่ 4.6 ค่าความแม่นยำ (Precision) ของการเลือกระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์หรือแผนศิลป์ มีค่าร้อยละ 82 และ100 ตามลำดับ ค่าโอกาสการเรียกซ้ำ (Recall) แผนการเรียนที่ 1 หรือ 2 ที่ร้อยละ 100 และ47 ตามลำดับ ผลที่ได้จากแบบจำลองด้วยเทคนิค Artificial Neural Network จะใช้การตรวจสอบ แบบ F1-score เพื่อตรวจสอบความแม่นยำของแบบจำลองอยู่ที่ ร้อยละ 90 และ64 จำนวนครั้งที่จะเกิดการเลือกแผนการเรียนที่ 1 และ2 (Support) รวมกันที่ 52 โดยค่าร้อยละความถูกต้อง (Accuracy) ของ Artificial Neural Network มีความถูกต้องที่ร้อยละ 85

1.7 การทดสอบแบบจำลองข้อมูลชุดที่ 2 ด้วยเทคนิค Naive Bayes

จากการทดลองใช้แบบจำลองด้วยเทคนิค Naive Bayes ใช้วิธีการลดขนาดข้อมูล (Principal Component Analysis : PCA) เมื่อแบ่งข้อมูลชุดเรียนรู้ (Training set) ร้อยละ 80 และข้อมูลชุดทดสอบ (Testing set) ร้อยละ 20 ค่าความแม่นยำ (Precision) ของการเลือกระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์ (SM) แทนด้วยเลข 1 หรือแผนศิลป์ (ART) แทนด้วยเลข 2

ตารางที่ 4.7 ผลการทดสอบแบบจำลองข้อมูลชุดที่ 2 ด้วยเทคนิค Naive Bayes

Naive Bayes	**PCA**			
	Precision	Recall	F1 - score	Support
SM 1	0.83	0.95	0.89	37
ART 2	0.80	0.53	0.64	15
Accuracy			0.83	52

จากตารางที่ 4.7 ค่าความแม่นยำ (Precision) ของการเลือกระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์หรือแผนศิลป์ มีค่าร้อยละ 83 และ80 ตามลำดับ ค่าโอกาสการเรียกซ้ำ (Recall) แผนการเรียนที่ 1 หรือ 2 ที่ร้อยละ 95 และ53 ตามลำดับ ผลที่ได้จากแบบจำลองด้วยเทคนิค Naive Bayes จะใช้การตรวจสอบ แบบ F1-score เพื่อตรวจสอบความแม่นยำของแบบจำลองอยู่ที่ร้อยละ 89 และ64 จำนวนครั้งที่จะเกิดการเลือกแผนการเรียนที่ 1 และ2 (Support) รวมกันที่ 52 โดยค่าร้อยละความถูกต้อง (Accuracy) ของ Naive Bayes มีความถูกต้องที่ร้อยละ 83

1.8 การทดสอบแบบจำลองข้อมูลชุดที่ 2 ด้วยเทคนิค Support Vector Machine

เมื่อใช้แบบจำลองด้วยเทคนิค Support Vector Machine ใช้วิธีการลดขนาดข้อมูล (Principal Component Analysis : PCA) โดยแบ่งข้อมูลชุดเรียนรู้ (Training Set) ร้อยละ 80 และข้อมูลชุดทดสอบ (Testing set) ร้อยละ 20 ค่าความแม่นยำ (Precision) ของการเลือกระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์ (SM) แทนด้วยเลข 1 หรือแผนศิลป์ (ART) แทนด้วยเลข 2

ตารางที่ 4.8 ผลการทดสอบแบบจำลองข้อมูลชุดที่ 2 ด้วยเทคนิค Support Vector Machine

Support Vector Machine	**PCA**			
	Precision	Recall	F1 - score	Support
SM 1	0.85	0.92	0.88	37
ART 2	0.75	0.60	0.67	15
Accuracy			0.83	52

จากตารางที่ 4.8 ค่าความแม่นยำ (Precision) ของการเลือกระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์หรือแผนศิลป์ มีค่าร้อยละ 85 และ75 ตามลำดับ ค่าโอกาสการเรียกซ้ำ (Recall) แผนการเรียนที่ 1 หรือ 2 ที่ร้อยละ 92 และ60 ตามลำดับ ผลที่ได้จากแบบจำลองด้วยเทคนิค Support Vector Machine จะใช้การตรวจสอบ แบบ F1-score เพื่อตรวจสอบความแม่นยำของแบบจำลองอยู่ที่ ร้อยละ 88 และ67 จำนวนครั้งที่จะเกิดการเลือกแผนการเรียนที่ 1 และ2 (Support) รวมกันที่ 52 โดยค่าร้อยละความถูกต้อง (Accuracy) ของ Support Vector Machine มีความถูกต้องที่ร้อยละ 83

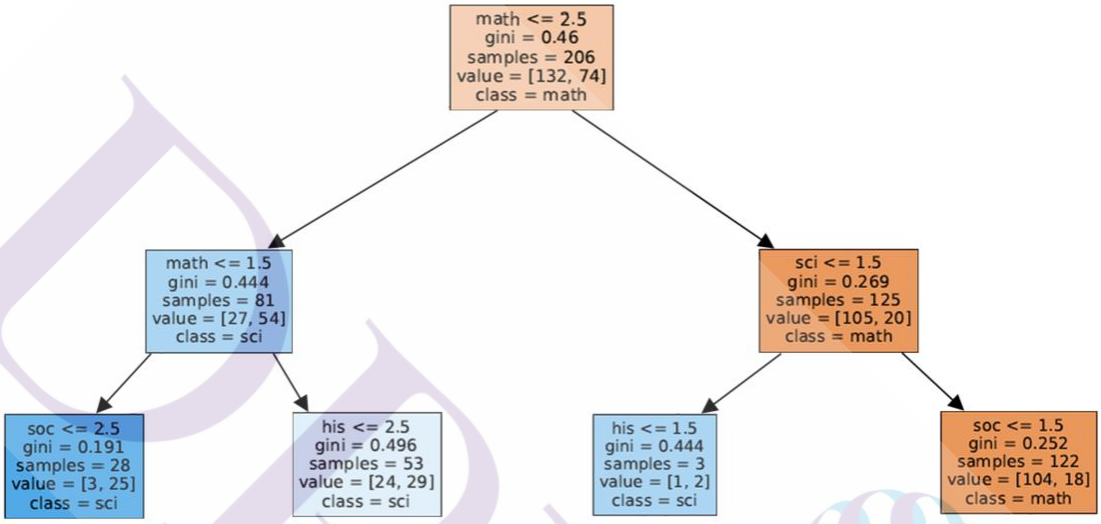
1.9 การทดสอบแบบจำลองชุดที่ 3 ด้วยเทคนิค Decision Tree

จากแบบจำลองด้วยเทคนิค Decision Tree กำหนดข้อมูลชุดที่ 3 โดยวิธี Backward Selection เริ่มต้นด้วยแบ่งข้อมูลชุดเรียนรู้ (Training Set) ร้อยละ 80 และข้อมูลชุดทดสอบ (Testing set) ร้อยละ 20 จากนั้นกำหนดค่าการเลือกระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์ (SM) ซึ่งแทนด้วยเลข 1 หรือแผนศิลป์ (ART) แทนด้วยเลข 2

ตารางที่ 4.9 ผลการทดสอบแบบจำลองชุดที่ 3 ด้วยเทคนิค Decision Tree

Decision Tree	**Backward Selection**			
	Precision	Recall	F1 - score	Support
SM 1	0.87	0.94	0.91	36
ART 2	0.85	0.69	0.76	16
Accuracy			0.87	52

จากตารางที่ 4.9 มีค่าความแม่นยำ (Precision) ของการเลือกระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์หรือแผนศิลป์ ร้อยละ 87 และ85 ตามลำดับ ค่าโอกาสการเรียกซ้ำ (Recall) แผนการเรียนที่ 1 และแผนการเรียนที่ 2 ที่ร้อยละ 94 และ69 ตามลำดับ ผลที่ได้จากแบบจำลองด้วยเทคนิค Decision Tree จะใช้การตรวจสอบ แบบ F1-score เพื่อตรวจสอบความแม่นยำของแบบจำลองอยู่ที่ ร้อยละ 91 และ76 จำนวนครั้งที่จะเกิดการเลือกแผนการเรียนที่ 1 และ2 (Support) รวมกันที่ 52 โดยค่าร้อยละความถูกต้อง (Accuracy) ของ Decision Tree มีความถูกต้องที่ร้อยละ 87



ภาพที่ 4.2 แผนผังต้นไม้จากการทำนายข้อมูลชุดที่ 3 ด้วยเทคนิค Decision Tree

จากภาพที่ 4.2 อธิบายลักษณะการทำนายผลข้อมูลชุดที่ 3 Backward Selection ด้วยเทคนิค Decision Tree จากข้อมูล 258 ชุด Machine Learning สามารถหา Root node ได้โดยค่า Gini ที่มีค่าสูงสุดในการทำนายนี้ เริ่มต้นจากคำถามแรกคือ ค่าเกรดเฉลี่ยของวิชาคณิตศาสตร์เป็นตัวเริ่มต้นแสดงให้เห็นว่าใน Attributes ทั้งหมด วิชาคณิตศาสตร์เป็นตัวจำแนกได้ดีที่สุดโดยทำนายได้ 206 จาก 258 ชุด และคำถามคือเกรดเฉลี่ยของวิชาคณิตศาสตร์ มากหรือน้อยกว่าโดยหากน้อยกว่า 2.5 ให้เลือกถามคำถามถัดไปที่ว่าคะแนนเฉลี่ยวิชาคณิตศาสตร์ด้านซ้ายมือมากหรือน้อยกว่า 1.5 แต่ถ้าคำถามแรกคะแนนวิชาคณิตศาสตร์มากกว่า 2.5 ให้ถามคำถามขวามือว่าคะแนนวิชาวิทยาศาสตร์มากกว่าหรือน้อยกว่า 1.5 เพื่อเลือกคำถามถัดไปจนจบสายของแผนผังต้นไม้จากการทำนายด้วยเทคนิค Decision Tree

1.10 การทดสอบแบบจำลองข้อมูลชุดที่ 3 ด้วยเทคนิค Artificial Neural Network

เมื่อทดลองแบบจำลองด้วยเทคนิค Artificial Neural Network กำหนดข้อมูลชุดที่ 3 โดยวิธี Backward Selection โดยแบ่งข้อมูลชุดเรียนรู้ (Training set) ร้อยละ 80 และข้อมูลชุดทดสอบ (Testing set) ร้อยละ 20 ค่าความแม่นยำ (Precision) ของการเลือกระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์ (SM) แทนด้วยเลข 1 หรือแผนศิลป์ (ART) แทนด้วยเลข 2

ตารางที่ 4.10 ผลการทดสอบแบบจำลองข้อมูลชุดที่ 3 ด้วยเทคนิค Artificial Neural Network

ANN	Backward Selection			
	Precision	Recall	F1 - score	Support
SM 1	0.91	0.84	0.87	37
ART 2	0.67	0.80	0.73	15
Accuracy			0.83	52

จากตารางที่ 4.10 ค่าความแม่นยำ (Precision) ของการเลือกระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์หรือแผนศิลป์ มีค่าร้อยละ 91 และ67 ตามลำดับ ค่าโอกาสการเรียกซ้ำ (Recall) แผนการเรียนที่ 1 หรือ 2 ที่ร้อยละ 84 และ80 ตามลำดับ ผลที่ได้จากแบบจำลองด้วยเทคนิค Artificial Neural Network จะใช้การตรวจสอบ แบบ F1-score เพื่อตรวจสอบความแม่นยำของแบบจำลองอยู่ที่ ร้อยละ 87 และ73 จำนวนครั้งที่จะเกิดการเลือกแผนการเรียนที่ 1 และ2 (Support) รวมกันที่ 52 โดยค่าร้อยละความถูกต้อง (Accuracy) ของ Artificial Neural Network มีความถูกต้องที่ร้อยละ 83

1.11 การทดสอบแบบจำลองข้อมูลชุดที่ 3 ด้วยเทคนิค Naive Bayes

จากการทดลองใช้แบบจำลองด้วยเทคนิค Naive Bayes กำหนดข้อมูลชุดที่ 3 โดยวิธี Backward Selection เมื่อแบ่งข้อมูลชุดเรียนรู้ (Training set) ร้อยละ 80 และข้อมูลชุดทดสอบ (Testing set) ร้อยละ 20 ค่าความแม่นยำ (Precision) ของการเลือกระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์ (SM) แทนด้วยเลข 1 หรือแผนศิลป์ (ART) แทนด้วยเลข 2

ตารางที่ 4.11 ผลการทดสอบแบบจำลองข้อมูลชุดที่ 3 ด้วยเทคนิค Naive Bayes

Naive Bayes	**Backward Selection**			
	Precision	Recall	F1 - score	Support
SM 1	0.90	0.95	0.92	37
ART 2	0.85	0.73	0.79	15
Accuracy			0.88	52

จากตารางที่ 4.11 ค่าความแม่นยำ (Precision) ของการเลือกระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์หรือแผนศิลป์ มีค่าร้อยละ 90 และ85 ตามลำดับ ค่าโอกาสการเรียกซ้ำ (Recall) แผนการเรียนที่ 1 หรือ 2 ที่ร้อยละ 95 และ73 ตามลำดับ ผลที่ได้จากแบบจำลองด้วยเทคนิค Naive Bayes จะใช้การตรวจสอบ แบบ F1-score เพื่อตรวจสอบความแม่นยำของแบบจำลองอยู่ที่ร้อยละ 92 และ79 จำนวนครั้งที่จะเกิดการเลือกแผนการเรียนที่ 1 และ2 (Support) รวมกันที่ 52 โดยค่าร้อยละความถูกต้อง (Accuracy) ของ Naive Bayes มีความถูกต้องที่ร้อยละ 88

1.12 การทดสอบแบบจำลองข้อมูลชุดที่ 3 ด้วยเทคนิค Support Vector Machine

เมื่อใช้แบบจำลองด้วยเทคนิค Support Vector Machine กำหนดข้อมูลชุดที่ 3 โดยวิธี Backward Selection โดยแบ่งข้อมูลชุดเรียนรู้ (Training Set) ร้อยละ 80 และข้อมูลชุดทดสอบ (Testing set) ร้อยละ 20 ค่าความแม่นยำ (Precision) ของการเลือกระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์ (SM) แทนด้วยเลข 1 หรือแผนศิลป์ (ART) แทนด้วยเลข 2

ตารางที่ 4.12 ผลการทดสอบแบบจำลองข้อมูลชุดที่ 3 ด้วยเทคนิค Support Vector Machine

Support Vector Machine	**Backward Selection**			
	Precision	Recall	F1 - score	Support
SM 1	0.91	0.86	0.89	37
ART 2	0.71	0.80	0.75	15
Accuracy			0.85	52

จากตารางที่ 4.12 ค่าความแม่นยำ (Precision) ของการเลือกระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์หรือแผนศิลป์ มีค่าร้อยละ 91 และ71 ตามลำดับ ค่าโอกาสการเรียกซ้ำ (Recall) แผนการเรียนที่ 1 หรือ 2 ที่ร้อยละ 86 และ 80 ตามลำดับ ผลที่ได้จากแบบจำลองด้วยเทคนิค Support Vector Machine จะใช้การตรวจสอบ แบบ F1-score เพื่อตรวจสอบความแม่นยำของแบบจำลองอยู่ที่ ร้อยละ 89 และ75 จำนวนครั้งที่จะเกิดการเลือกแผนการเรียนที่ 1 และ2 (Support) รวมกันที่ 52 โดยค่าร้อยละความถูกต้อง (Accuracy) ของ Support Vector Machine มีความถูกต้องที่ร้อยละ 85

- ตอนที่ 2 การเปรียบเทียบผลการวิเคราะห์ประสิทธิภาพโมเดลด้วยข้อมูลชุดที่ 1 ชุดที่ 2 และชุดที่ 3

ตารางที่ 4.13 การเปรียบเทียบเทคนิค Decision Tree ด้วยข้อมูลชุดที่ 1 ชุดที่ 2 และชุดที่ 3

Decision Tree	**Interpretation Evaluation**			
	Precision	Recall	F1 - score	Support
ข้อมูลชุดที่ 1 (Correlation Analysis)				
SM 1	0.82	0.89	0.85	36
ART 2	0.69	0.59	0.62	16
Accuracy			0.79	52
ข้อมูลชุดที่ 2 (PCA)				
SM 1	0.80	0.92	0.86	36
ART 2	0.73	0.50	0.59	16
Accuracy			0.79	52
ข้อมูลชุดที่ 3 (Backward Selection)				
SM 1	0.87	0.94	0.91	36
ART 2	0.85	0.69	0.76	16
Accuracy			0.87	52

จากตารางที่ 4.13 ผลการวิเคราะห์ประสิทธิภาพโมเดลการเลือกแผนการเรียนวิทยาศาสตร์และคณิตศาสตร์หรือแผนศิลป์ ของข้อมูลชุดที่ 1 และชุดที่ 2 มีค่าร้อยละความถูกต้อง (Accuracy) ของโมเดล Decision Tree เท่ากันทั้งสองวิธี อาจเนื่องมาจากการนำข้อมูลมาผ่านการวิเคราะห์ในกระบวนการลดรูป PCA ของข้อมูลชุดที่ 2 เกิดจากการสร้างองค์ประกอบใหม่ภายใต้ชุดข้อมูลที่ 1 ซึ่งส่งผลทำให้องค์ประกอบของโครงสร้างเดิมถูกรวมและลดจำนวนลงไปสร้างเป็นองค์ประกอบใหม่ที่จำนวนน้อยลงจากข้อมูลชุดที่ 1 ประกอบด้วย 12 Attributes ผ่านข้อมูลชุดที่ 2 PCA รวมองค์ประกอบใหม่เป็น 4 Components ทำให้ค่าร้อยละความถูกต้อง (Accuracy) ของข้อมูลชุดที่ 1 และ2 เท่ากันที่ร้อยละ 79

โดยมีการรายงานค่าความแม่นยำ (Precision) การเลือกระหว่างแผนการเรียนที่ 1 และ 2 ของข้อมูลชุดที่ 1 มีค่าร้อยละ 82 และ69 ตามลำดับ ข้อมูลชุดที่ 2 มีค่าร้อยละ 80 และ73 ตามลำดับ ค่าโอกาสการเรียกซ้ำ (Recall) แผนการเรียนที่ 1 หรือ 2 ของข้อมูลชุดที่ 1 มีค่าร้อยละ 89 และ59 ตามลำดับ ของข้อมูลชุดที่ 2 มีค่าร้อยละ 92 และ50 ตามลำดับ ผลที่ได้จากแบบจำลองด้วยเทคนิค Decision Tree จะใช้การตรวจสอบ แบบ F1-score เพื่อตรวจสอบความแม่นยำของแบบจำลองการเลือกแผนการเรียนที่ 1 หรือ 2 ของข้อมูลชุดที่ 1 อยู่ที่ ร้อยละ 85 และ62 ตามลำดับ ข้อมูลชุดที่ 2 อยู่ที่ ร้อยละ 86 และ59 โดยจำนวนครั้งที่จะเกิดการเลือกแผนการเรียนที่ 1 และ2 (Support) รวมกันที่ 52 ทั้งข้อมูลชุดที่ 1 และ 2

ในกรณีข้อมูลชุดที่ 3 ใช้วิธีการ Backward Selection เป็นการคัดเลือก Attributes โดยการนำ Attributes ทุกตัวเข้ามาในสมการและดำเนินการพิจารณา Attributes ที่มีค่าสัมประสิทธิ์สหสัมพันธ์บางส่วน (Partial Correlation) โดยโมเดลวิเคราะห์แผนการเรียนด้วย วิธี Decision Tree ที่ใช้ Backward Selection ได้เลือก Attributes (math, sci, soc, his, pe) มีค่าร้อยละความถูกต้อง (Accuracy) ร้อยละ 87 โดยมีการรายงานค่าความแม่นยำ (Precision) ของการเลือกระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์หรือแผนศิลป์ ร้อยละ 87 และ85 ตามลำดับ ค่าโอกาสการเรียกซ้ำ (Recall) แผนการเรียนที่ 1 และแผนการเรียนที่ 2 ที่ร้อยละ 94 และ69 ตามลำดับ ผลที่ได้จากแบบจำลองด้วยเทคนิค Decision Tree จะใช้การตรวจสอบ แบบ F1-score เพื่อตรวจสอบความแม่นยำของแบบจำลองอยู่ที่ ร้อยละ 91 และ76 จำนวนครั้งที่จะเกิดการเลือกแผนการเรียนที่ 1 และ2 (Support) รวมกันที่ 52

ตารางที่ 4.14 การเปรียบเทียบเทคนิค Artificial Neural Network ด้วยข้อมูลชุดที่ 1 ชุดที่ 2 และชุดที่ 3

ANN	Interpretation Evaluation			
	Precision	Recall	F1 - score	Support
ข้อมูลชุดที่ 1 (Correlation Analysis)				
SM 1	0.87	0.89	0.88	37
ART 2	0.71	0.67	0.69	15
Accuracy			0.83	91
ข้อมูลชุดที่ 2 (PCA)				
SM 1	0.82	1.00	0.90	37
ART 2	1.00	0.47	0.64	15
Accuracy			0.85	52
ข้อมูลชุดที่ 3 (Backward Selection)				
SM 1	0.91	0.84	0.87	37
ART 2	0.67	0.80	0.73	15
Accuracy			0.83	52

จากตารางที่ 4.14 ผลการวิเคราะห์ประสิทธิภาพโมเดลการเลือกแผนการเรียนวิทยาศาสตร์และคณิตศาสตร์หรือแผนศิลป์ ของข้อมูลชุดที่ 1 และชุดที่ 2 มีค่าร้อยละความถูกต้อง (Accuracy) ของโมเดล Artificial Neural Network ใกล้เคียงกันทั้งสองวิธีค่าความแต่ต่างกันอยู่ที่ร้อยละ 2 อาจเนื่องมาจากการนำข้อมูลมาผ่านการวิเคราะห์ในกระบวนการลดรูป PCA ของข้อมูลชุดที่ 2 เกิดจากการสร้างองค์ประกอบใหม่ภายใต้ชุดข้อมูลที่ 1 ซึ่งส่งผลทำให้องค์ประกอบของโครงสร้างเดิมถูกรวม และสร้างเป็นองค์ประกอบใหม่ขึ้นที่จำนวนน้อยลงจากข้อมูลชุดที่ 1 ประกอบด้วย 12 Attributes ผ่านข้อมูลชุดที่ 2 PCA รวมองค์ประกอบใหม่เป็น 4 Components ค่าร้อยละความถูกต้อง (Accuracy) ข้อมูลชุดที่ 1 มีค่าร้อยละ 83 และข้อมูลชุดที่ 2 มีค่า ร้อยละ 85

โดยมีการรายงานค่าความแม่นยำ (Precision) การเลือกระหว่างแผนการเรียนที่ 1 และ2 ของข้อมูลชุดที่ 1 มีค่าร้อยละ 87 และ71 ตามลำดับ ข้อมูลชุดที่ 2 มีค่าร้อยละ 82 และ100 ตามลำดับ ค่าโอกาสการเรียกซ้ำ (Recall) แผนการเรียนที่ 1 หรือ 2 ของข้อมูลชุดที่ 1 มีค่าร้อยละ 89 และ 67 ตามลำดับ ของข้อมูลชุดที่ 2 มีค่าร้อยละ 100 และ 47 ตามลำดับ ผลที่ได้จากแบบจำลองด้วยเทคนิค Artificial Neural Network จะใช้การตรวจสอบ แบบ F1-score เพื่อตรวจสอบความแม่นยำของ

แบบจำลองการเลือกแผนการเรียนที่ 1 หรือ 2 ของข้อมูลชุดที่ 1 อยู่ที่ ร้อยละ 88 และ69 ตามลำดับ ของข้อมูลชุดที่ 2 อยู่ที่ ร้อยละ 90 และ64 โดยจำนวนครั้งที่จะเกิดการเลือกแผนการเรียนที่ 1 และ2 (Support) รวมกันที่ 52 ทั้งข้อมูลชุดที่ 1 และ2

ในกรณีข้อมูลชุดที่ 3 ใช้วิธีการ Backward Selection เป็นการคัดเลือก Attributes โดยการนำ Attributes ทุกตัวเข้ามาในสมการและดำเนินการพิจารณา Attributes ที่มีค่าสัมประสิทธิ์สหสัมพันธ์บางส่วน (Partial Correlation) โดยโมเดลวิเคราะห์แผนการเรียนด้วยวิธี Artificial Neural Network ที่ใช้ Backward Selection ได้เลือก Attributes (math, eng, soc, heal, his) มีค่าร้อยละความถูกต้อง (Accuracy) ร้อยละ 83 โดยมีการรายงานค่าความแม่นยำ (Precision) ของการเลือกระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์หรือแผนศิลป์ ร้อยละ 91 และ67 ตามลำดับ ค่าโอกาสการเรียกซ้ำ (Recall) แผนการเรียนที่ 1 และแผนการเรียนที่ 2 ที่ร้อยละ 84 และ80 ตามลำดับ ผลที่ได้จากแบบจำลองด้วยเทคนิค Artificial Neural Network จะใช้การตรวจสอบ แบบ F1-score เพื่อตรวจสอบความแม่นยำของแบบจำลองอยู่ที่ ร้อยละ 87 และ73 จำนวนครั้งที่จะเกิดการเลือกแผนการเรียนที่ 1 และ2 (Support) รวมกันที่ 52

ตารางที่ 4.15 การเปรียบเทียบเทคนิค Naive Bayes ด้วยข้อมูลชุดที่ 1 ชุดที่ 2 และชุดที่ 3

Naive Bayes	**Interpretation Evaluation**			
	Precision	Recall	F1 - score	Support
ข้อมูลชุดที่ 1 (Correlation Analysis)				
SM 1	0.88	0.95	0.91	37
ART 2	0.83	0.67	0.74	15
Accuracy			0.87	52
ข้อมูลชุดที่ 2 (PCA)				
SM 1	0.83	0.95	0.89	37
ART 2	0.80	0.53	0.64	15
Accuracy			0.83	52
ข้อมูลชุดที่ 3 (Backward Selection)				
SM 1	0.90	0.95	0.92	37
ART 2	0.85	0.73	0.79	15
Accuracy			0.88	52

จากตารางที่ 4.15 ผลการวิเคราะห์ประสิทธิภาพโมเดลการเลือกแผนการเรียนวิทยาศาสตร์และคณิตศาสตร์หรือแผนศิลป์ ของข้อมูลชุดที่ 1 และชุดที่ 2 มีค่าร้อยละความถูกต้อง (Accuracy) ของโมเดล Naive Bayes ใกล้เคียงกันทั้งสองวิธีค่าความแตกต่างกันอยู่ที่ร้อยละ 4 อาจเนื่องมาจากการนำข้อมูลมาผ่านการวิเคราะห์ในกระบวนการลดรูป PCA ของข้อมูลชุดที่ 2 เกิดจากการสร้างองค์ประกอบใหม่ภายใต้ชุดข้อมูลที่ 1 ซึ่งส่งผลทำให้องค์ประกอบของโครงสร้างเดิมถูกรวมและสร้างเป็นองค์ประกอบใหม่ขึ้นที่จำนวนน้อยลงจากข้อมูลชุดที่ 1 ประกอบด้วย 12 Attributes ผ่านข้อมูลชุดที่ 2 PCA รวมองค์ประกอบใหม่เป็น 4 Components ค่าร้อยละความถูกต้อง (Accuracy) ข้อมูลชุดที่ 1 มีค่าร้อยละ 87 และข้อมูลชุดที่ 2 มีค่า ร้อยละ 83

โดยมีการรายงานค่าความแม่นยำ (Precision) การเลือกระหว่างแผนการเรียนที่ 1 และ2 ของข้อมูลชุดที่ 1 มีค่าร้อยละ 88 และ83 ตามลำดับ ข้อมูลชุดที่ 2 มีค่าร้อยละ 83 และ80 ตามลำดับ ค่าโอกาสการเรียกซ้ำ (Recall) แผนการเรียนที่ 1 หรือ 2 ของข้อมูลชุดที่ 1 มีค่าร้อยละ 95 และ67 ตามลำดับ ของข้อมูลชุดที่ 2 มีค่าร้อยละ 95 และ53 ตามลำดับ ผลที่ได้จากแบบจำลองด้วยเทคนิค Naive Bayes จะใช้การตรวจสอบ แบบ F1-score เพื่อตรวจสอบความแม่นยำของแบบจำลองการเลือกแผนการเรียนที่ 1 หรือ 2 ของข้อมูลชุดที่ 1 อยู่ที่ ร้อยละ 91 และ74 ตามลำดับของข้อมูลชุดที่ 2 อยู่ที่ ร้อยละ 89 และ64 โดยจำนวนครั้งที่จะเกิดการเลือกแผนการเรียนที่ 1 และ2 (Support) รวมกันที่ 52 ทั้งข้อมูลชุดที่ 1 และ2

ในกรณีข้อมูลชุดที่ 3 ใช้วิธีการ Backward Selection เป็นการคัดเลือก Attributes โดยการนำ Attributes ทุกตัวเข้ามาในสมการและดำเนินการพิจารณา Attributes ที่มีค่าสัมประสิทธิ์สหสัมพันธ์บางส่วน (Partial Correlation) โดยโมเดลวิเคราะห์แผนการเรียนด้วย วิธี Naive Bayes ที่ใช้ Backward Selection ได้เลือก Attributes (sex, math, soc, his, gpax) มีค่าร้อยละความถูกต้อง (Accuracy) ร้อยละ 88 โดยมีการรายงานค่าความแม่นยำ (Precision) ของการเลือกระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์หรือแผนศิลป์ ร้อยละ 90 และ85 ตามลำดับ ค่าโอกาสการเรียกซ้ำ (Recall) แผนการเรียนที่ 1 และแผนการเรียนที่ 2 ร้อยละ 95 และ73 ตามลำดับ ผลที่ได้จากแบบจำลองด้วยเทคนิค Naive Bayes จะใช้การตรวจสอบ แบบ F1-score เพื่อตรวจสอบความแม่นยำของแบบจำลองอยู่ที่ ร้อยละ 92 และ79 จำนวนครั้งที่จะเกิดการเลือกแผนการเรียนที่ 1 และ2 (Support) รวมกันที่ 52

ตารางที่ 4.16 การเปรียบเทียบเทคนิค Support Vector Machine ด้วยข้อมูลชุดที่ 1 ชุดที่ 2 และชุดที่ 3

SVM	Interpretation Evaluation			
	Precision	Recall	F1 - score	Support
ข้อมูลชุดที่ 1 (Correlation Analysis)				
SM 1	0.86	0.94	0.90	34
ART 2	0.87	0.72	0.79	18
Accuracy			0.87	52
ข้อมูลชุดที่ 2 (PCA)				
SM 1	0.85	0.92	0.88	37
ART 2	0.75	0.60	0.67	15
Accuracy			0.83	52
ข้อมูลชุดที่ 3 (Backward Selection)				
SM 1	0.91	0.86	0.89	37
ART 2	0.71	0.80	0.75	15
Accuracy			0.85	52

จากตารางที่ 4.16 ผลการวิเคราะห์ประสิทธิภาพโมเดลการเลือกแผนการเรียนวิทยาศาสตร์และคณิตศาสตร์หรือแผนศิลป์ ของข้อมูลชุดที่ 1 และชุดที่ 2 มีค่าร้อยละความถูกต้อง (Accuracy) ของโมเดล Support Vector Machine ใกล้เคียงกันทั้งสองวิธีค่าความแตกต่างกันอยู่ที่ร้อยละ 4 อาจเนื่องมาจากการนำข้อมูลมาผ่านการวิเคราะห์ในกระบวนการลดรูป PCA ของข้อมูลชุดที่ 2 เกิดจากการสร้างองค์ประกอบใหม่ภายใต้ชุดข้อมูลที่ 1 ซึ่งส่งผลทำให้องค์ประกอบของโครงสร้างเดิมถูกรวมและสร้างเป็นองค์ประกอบใหม่ขึ้นที่จำนวนน้อยลงจากข้อมูลชุดที่ 1 ประกอบด้วย 12 Attributes ผ่านข้อมูลชุดที่ 2 PCA รวมองค์ประกอบใหม่เป็น 4 Components ค่าร้อยละความถูกต้อง (Accuracy) ข้อมูลชุดที่ 1 มีค่าร้อยละ 87 และข้อมูลชุดที่ 2 มีค่าร้อยละ 83

โดยมีการรายงานค่าความแม่นยำ (Precision) การเลือกระหว่างแผนการเรียนที่ 1 และ2 ของข้อมูลชุดที่ 1 มีค่าร้อยละ 86 และ87 ตามลำดับ ข้อมูลชุดที่ 2 มีค่าร้อยละ 85 และ75 ตามลำดับ ค่าโอกาสการเรียกซ้ำ (Recall) แผนการเรียนที่ 1 หรือ 2 ของข้อมูลชุดที่ 1 มีค่าร้อยละ 94 และ72 ตามลำดับ ของข้อมูลชุดที่ 2 มีค่าร้อยละ 92 และ60 ตามลำดับ ผลที่ได้จากแบบจำลองด้วยเทคนิค Support Vector Machine จะใช้การตรวจสอบ แบบ F1-score เพื่อตรวจสอบความแม่นยำของ

แบบจำลองการเลือกแผนการเรียนที่ 1 หรือ 2 ของข้อมูลชุดที่ 1 อยู่ที่ ร้อยละ 90 และ79 ตามลำดับ ของข้อมูลชุดที่ 2 อยู่ที่ ร้อยละ 88 และ67 โดยจำนวนครั้งที่จะเกิดการเลือกแผนการเรียนที่ 1 และ2 (Support) รวมกันที่ 52 ทั้งข้อมูลชุดที่ 1 และ 2

ในกรณีข้อมูลชุดที่ 3 ใช้วิธีการ Backward Selection เป็นการคัดเลือก Attributes โดยการนำ Attributes ทุกตัวเข้ามาในสมการและดำเนินการพิจารณา Attributes ที่มีค่าสัมประสิทธิ์สหสัมพันธ์บางส่วน (Partial Correlation) โดยโมเดลวิเคราะห์แผนการเรียนด้วย วิธี Support Vector Machine ที่ใช้ Backward Selection ได้เลือก Attributes (math, sci, eng, pe, gpax) มีค่าร้อยละความถูกต้อง (Accuracy) ร้อยละ 85 โดยมีการรายงานค่าความแม่นยำ (Precision) ของการเลือกระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์หรือแผนศิลป์ ร้อยละ 91 และ71 ตามลำดับ ค่าโอกาสการเรียกซ้ำ (Recall) แผนการเรียนที่ 1 และแผนการเรียนที่ 2 ที่ร้อยละ 86 และ80 ตามลำดับ ผลที่ได้จากแบบจำลองด้วยเทคนิค Support Vector Machine จะใช้การตรวจสอบ แบบ F1-score เพื่อตรวจสอบความแม่นยำของแบบจำลองอยู่ที่ ร้อยละ 89 และ75 จำนวนครั้งที่จะเกิดการเลือกแผนการเรียนที่ 1 และ2 (Support) รวมกันที่ 52

- ตอนที่ 3 สรุปผลการวิเคราะห์

ตารางที่ 4.17 สรุปผลการวิเคราะห์ค่า Accuracy ด้วยข้อมูลชุดที่ 1 ข้อมูลชุดที่ 2 และข้อมูลชุดที่ 3

Algorithms	Accuracy ข้อมูลชุดที่ 1 (Correlation Analysis)	Accuracy ข้อมูลชุดที่ 2 (PCA)	Accuracy ข้อมูลชุดที่ 3 (Backward Selection)
Decision Tree	79	79	87
Artificial Neural Network	83	85	83
Naive Bayes	87	83	88
Support Vector Machine	87	83	85

จากตารางที่ 4.17 พบว่า ข้อมูลชุดที่ 1 มีค่าเฉลี่ยร้อยละความถูกต้อง (Accuracy) เมื่อเปรียบเทียบกับข้อมูลชุดที่ 2 มีความแตกต่างกันไม่มากเมื่อเปรียบเทียบรายวิธี เริ่มจากเทคนิค Support Vector Machine และNaive Bayes มีค่าร้อยละความถูกต้อง (Accuracy) สูงที่สุดเท่ากันที่ร้อยละ 87 ลำดับต่อมาเป็นเทคนิค Artificial Neural Network ร้อยละ 83 ลำดับสุดท้ายเป็นเทคนิค Decision Tree ร้อยละ 79 และข้อมูลชุดที่ 2 มีค่าเฉลี่ยร้อยละความถูกต้อง (Accuracy) ของเทคนิค Artificial Neural Network สูงที่สุด ร้อยละ 85 ลำดับที่สองเป็นเทคนิค Naive Bayes และSupport Vector Machine เท่ากันที่ร้อยละ 83 ลำดับสุดท้ายเป็นเทคนิค Decision Tree ร้อยละ 79 ค่าเฉลี่ยร้อยละความถูกต้อง (Accuracy) ของข้อมูลชุดที่ 3 มีค่าสูงสุดจากเทคนิค Naive Bayes อยู่ที่ร้อยละ 88 ลำดับที่สองมาเป็นเทคนิค Decision Tree ร้อยละ 87 ลำดับที่สามมาเป็นเทคนิค Support Vector Machine ร้อยละ 85 ลำดับสุดท้ายเป็นเทคนิค Artificial Neural Network ร้อยละ 83

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

จากการดำเนินการวิจัยแนวโน้มการเลือกแผนการเรียนมัธยมศึกษาตอนปลายด้วยการใช้ Machine Learning ประกอบด้วย 1) Decision Tree 2) Artificial Neural Network 3) Naive Bayes 4) Support Vector Machine นำข้อมูลเกรดเฉลี่ยของผู้สำเร็จการศึกษาชั้นมัธยมศึกษาปีที่ 4 - 6 ปีการศึกษา 2554 - 2556 มาทำการเลือกผู้มีผลการศึกษาเกรดเฉลี่ยระดับ 3.00 ขึ้นไปเพื่อนำมาใช้กับ Machine Learning ผลการคัดเลือกจาก 454 ได้จำนวนข้อมูล 258 ชุด ในการวิจัยครั้งนี้เริ่มจากข้อมูลที่ใช้ในการสร้างและหาประสิทธิภาพแบบจำลองทั้ง 4 แบบ และทดลองด้วยข้อมูล 3 ชุด ดังนี้

1) ข้อมูลชุดที่ 1 เป็นชุดข้อมูล ที่ผ่านกระบวนการเตรียมข้อมูล (Data Preparation) ด้วย Correlation Analysis มีข้อมูลจำนวน 12 Attributes

2) ข้อมูลชุดที่ 2 เป็นชุดข้อมูลที่ใช้วิธีการลดขนาดข้อมูล (Principal Component Analysis : PCA) มีข้อมูลจำนวน 4 Components

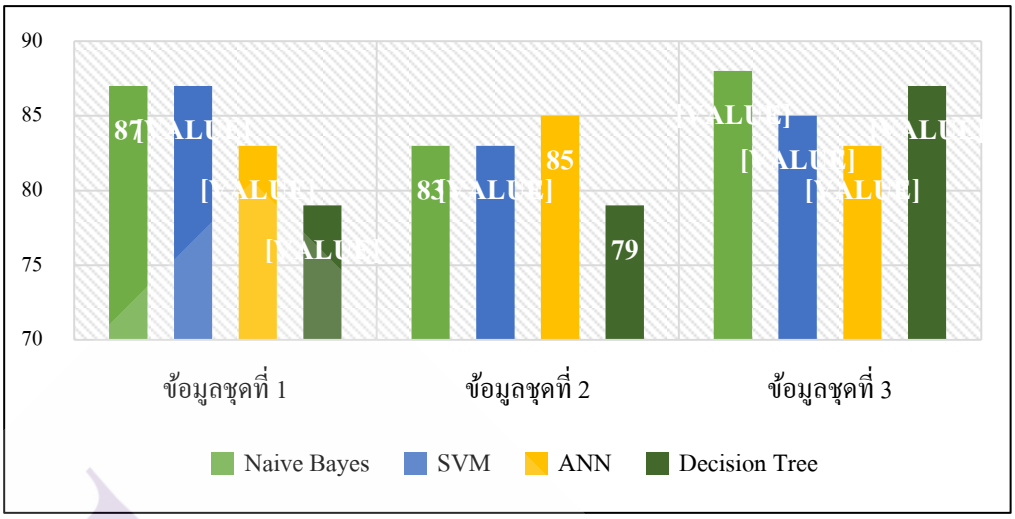
3) ข้อมูลชุดที่ 3 เป็นชุดข้อมูลที่ใช้วิธีการวิเคราะห์ตัวเลือกคุณลักษณะตามลำดับ Feature Selector (Sequential Feature Selector : Backward Selection) โดยจะมีการเลือก Attributes ที่แตกต่างกัน ดังนี้

3.1) Decision Tree เลือก Attributes (math, sci, soc, his, pe)

3.2) Artificial Neural Network เลือก Attributes (math, eng, soc, heal, his)

3.3) Naive Bayes เลือก Attributes (sex, math, soc, his, gpax)

3.4) Support Vector Machine เลือก Attributes (math, sci, eng, pe, gpax)



ภาพที่ 5.1 การเปรียบเทียบค่าความถูกต้อง (Accuracy) ด้วยข้อมูลชุดที่ 1 ข้อมูลชุดที่ 2 และ ข้อมูลชุดที่ 3

จากภาพที่ 5.1 ข้อมูลชุดที่ 1 ใช้วิธี Correlation Analysis มีข้อมูลจำนวน 12 Attributes ค่าเฉลี่ยร้อยละความถูกต้อง (Accuracy) เรียงลำดับจากมากไปหาน้อยดังนี้ Naive Bayes และ Support Vector Machine มีค่าเท่ากันที่ ร้อยละ 87 ลำดับต่อมา Artificial Neural Network ร้อยละ 83 และลำดับสุดท้าย Decision Tree ร้อยละ 79 ข้อมูลชุดที่ 2 ใช้วิธีการลดขนาดข้อมูล (Principal Component Analysis : PCA) มีข้อมูลจำนวน 4 Components ค่าเฉลี่ยร้อยละความถูกต้อง (Accuracy) เรียงลำดับจากมากไปหาน้อยดังนี้ Artificial Neural Network ร้อยละ 85 ลำดับต่อมา Naive Bayes และ Support Vector Machine เท่ากันที่ร้อยละ 83 และลำดับสุดท้าย Decision Tree ร้อยละ 79 ข้อมูลชุดที่ 3 ใช้วิธีการวิเคราะห์ตัวเลือกคุณลักษณะตามลำดับ Feature Selector (Sequential Feature Selector : Backward Selection) ค่าเฉลี่ยร้อยละความถูกต้อง (Accuracy) เรียงลำดับจากมากไปหาน้อยดังนี้ Naive Bayes ร้อยละ 88 ลำดับต่อมา Decision Tree ร้อยละ 87 ลำดับต่อมา Support Vector Machine ร้อยละ 85 และลำดับสุดท้าย Artificial Neural Network ร้อยละ 83

5.1 อภิปรายผล

การวิจัยครั้งนี้ใช้ Machine Learning ร่วมกับ Scikit - Learn Package มาวิเคราะห์ข้อมูลผู้เรียนแบบ Supervised Learning เปรียบเทียบ 4 วิธี และเตรียมชุดข้อมูลที่ผ่านกระบวนการ (Data Preparation) ที่แตกต่างกัน 3 ชุด เมื่อพิจารณาผลจากข้อมูลชุดที่ 1 ชุดที่ 2 และชุดที่ 3 จะเห็นว่าค่าเฉลี่ยร้อยละความถูกต้อง (Accuracy) ไม่ได้แต่ต่างกันมาก อาจเนื่องมาจากมีกระบวนการเตรียมข้อมูลแม้ว่าจะเป็นวิธีการที่แตกต่างกัน ซึ่งอาจส่งผลต่อค่าเฉลี่ยร้อยละความถูกต้อง (Accuracy) ในแต่ละเทคนิคบ้าง ดังนั้นในการเลือกใช้ชุดข้อมูลที่มีการเตรียมข้อมูลแตกต่างกัน อาจพิจารณาได้จากผลของอัลกอริทึมที่เราสนใจจะเลือกใช้ โดยข้อมูลชุดที่ 1 มีลักษณะเด่นในเรื่องการลด Attributes ลงจากข้อมูลทั้งหมดจำนวน 70 Attributes เหลือจำนวน 12 Attributes ซึ่งจำนวนที่คงอยู่เป็นส่วนที่เห็นถึงความสัมพันธ์ ที่ผ่านการทำ Correlation Analysis จึงยังคงมีตัวแปรมากที่สุดเมื่อเปรียบเทียบกับอีกสองวิธี โดยอัลกอริทึมที่ต้องอาศัยการอธิบายความสัมพันธ์ระหว่างข้อมูลอย่าง Decision Tree จะมีความเหมาะสมสามารถนำไปใช้ช่วยในการแนะแนวหาตำตอบได้จากการสร้างแผนภูมิต้นไม้ สำหรับครูแนะแนวโดยการนำเอาแผนภูมิต้นไม้นี้ไปใช้ในการตั้งคำถามตามลักษณะของการแตกตัวของโหนดต่างๆ ทำให้นักเรียนเห็นแนวโน้มช่วยในการตัดสินใจเลือกแผนการเรียนได้ และเมื่อพิจารณาผลของอัลกอริทึมอีกสามวิธี ข้อมูลชุดที่ 1 ก็ได้ผลเป็นที่น่าพอใจในค่าเฉลี่ยร้อยละความถูกต้อง (Accuracy) ข้อมูลชุดที่ 2 มีลักษณะเด่นในเรื่องการลด Attributes ลงโดยเกิดจากการวิเคราะห์ความแปรปรวนด้วยการสร้างตัว Object ขึ้นมาโดยบอกจำนวน Components ในที่นี้คือ x_t หมายถึง ข้อมูลที่ Transform มาแล้ว หลังจากนั้น PCA ก็จะส่งค่ากลับมาใน Scikit-learn เรียกว่า Explained variance หรือ Eigenvector ในการสร้าง Components ดังนั้นจึงเป็นวิธีที่เหมาะสำหรับการจะลดตัวแปรลงด้วยการรวมและสร้างขึ้นมาใหม่โดยไม่เสียคุณสมบัติไปแต่จะอธิบายแผนผังได้ยากเนื่องจากได้รวมเป็น Components ใหม่แล้วแยกดูองค์ประกอบภายในได้ยาก ข้อมูลชุดที่ 3 เป็นวิธีที่มีค่าเฉลี่ยร้อยละความถูกต้อง (Accuracy) สูงในทุกอัลกอริทึมที่ใช้สร้าง Machine Learning โดยดำเนินการพิจารณา Attributes ที่มีค่าสัมประสิทธิ์สหสัมพันธ์บางส่วน (Partial Correlation) ทำให้การเลือก Attributes ในแต่ละอัลกอริทึมจะมีความแตกต่างกัน

ดังนั้นในการสร้าง Machine Learning การเลือกใช้อัลกอริทึมอาจขึ้นอยู่กันความคุ้มค่าทางด้านเวลาที่ใช้ในการคำนวน จำนวนปริมาณข้อมูล การประมวลผล และการอธิบายความสัมพันธ์ระหว่างข้อมูลเทคนิค Artificial Neural Network เป็นหนึ่งในวิธีที่มีประสิทธิภาพสูงแต่ต้องแลกมาด้วยการใช้ทรัพยากรในการประมวลผลและใช้เวลาในการคำนวนนานกว่าวิธีอื่นในการวิจัยครั้งนี้ เทคนิค Support Vector Machine และ Naive Bayes มีค่าความถูกต้อง (Accuracy) สูงอาจเนื่องมาจาก Support Vector Machine [8]และNaive Bayes เป็นวิธีที่เหมาะสมกับชุดข้อมูลที่

ผู้วิจัยใช้โดยอัลกอริทึมนี้มีความสามารถแยกและทำนายข้อมูล เป็นการหาระยะห่างระหว่างกลุ่มข้อมูลที่ต้องการแยก (Hyper plane) ที่ใกล้ที่สุด นอกจากนี้หากพิจารณา จากค่าความแม่นยำ (Precision) ของการเลือกระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์ซึ่งแทนด้วยเลข 1 หรือแผนศิลป์แทนด้วยเลข 2 และค่าโอกาสการเรียกซ้ำ (Recall) จะพบว่ามีค่าสูง ทั้งสองแผนการเรียน จึงน่าเชื่อได้ว่ามีความเหมาะสมในการใช้หาแนวโน้มการศึกษาต่อมัธยมศึกษาตอนปลาย ซึ่งสอดคล้องกับ นนท์ บุญนิธิประเสริฐ และชัยพร เขมะภาตะพันธ์ ที่ได้นำเสนอวิธีการกรองสแปมในระบบส่งข้อความของโทรศัพท์เคลื่อนที่ ด้วยวิธีการกรองโดยใช้ Support Vector Machine และ Naive Bayes หลักการของการกรองสแปม ในงานวิจัยจะมีการแบ่งคลาสเป็น 2 แบบ คือใช่ และไม่ใช่ ใช้ลักษณะเป็นเหมือนกันกับการเลือกระหว่างแผนวิทยาศาสตร์และคณิตศาสตร์ซึ่งแทนด้วยเลข 1 หรือแผนศิลป์แทนด้วยเลข 2 โดยผลทดสอบด้วยวิธีการกรองแบบ Support Vector Machine มีความถูกต้องในการกรองข้อมูลสูงกว่า Naive Bayes แต่วิธีการกรองแบบ Naive Bayes ใช้เวลาในการประมวลผลน้อยกว่า

5.2 ปัญหาและข้อเสนอแนะ

5.2.1 ปัญหาและข้อจำกัดที่พบในงานวิจัย

5.2.1.1 ข้อมูลที่ใช้ในการวิจัยเป็นการรวบรวมข้อมูลที่มีอยู่เดิมในระบบถึงแม้จะรวมรวมมาทุกขอบเขตข้อมูล แต่ก็อาจไม่ครอบคลุมมากพอที่จะทำให้ผลของการทำนายมากขึ้นกว่าที่เป็นแม้จะพยายามใช้กระบวนการต่างๆ เข้ามาช่วยในการจะพัฒนาแบบจำลองในหลากหลายเทคนิคก็ตาม

5.2.1.2 ในการใช้แบบจำลองที่มีเทคนิคซับซ้อนนั้นถึงแม้จะได้ค่าประสิทธิภาพสูงกว่าการใช้แบบจำลองพื้นฐานแต่ก็มีข้อสังเกต ที่ต้องพิจารณาเรื่องประสิทธิภาพค่าความถูกต้อง (Accuracy) ที่สูงแต่ต้องแลกมาด้วยความคุ้มค่าทางด้านเวลาหรือต้องใช้เครื่องมือในการคำนวนที่สูงตาม

5.2.2 ข้อเสนอแนะ

5.2.2.1 การวิจัยครั้งต่อไปควรมีการออกแบบฐานข้อมูลในการเก็บ เพื่อการทดลองตั้งแต่เริ่มต้นถึงแม้ต้องใช้เวลานานแต่ก็จะทำให้ได้ค่าประสิทธิภาพค่าความถูกต้อง (Accuracy) สูงขึ้น

5.2.2.2 ควรมีการพัฒนาฐานข้อมูลและการวิเคราะห์บนระบบ Cloud computing เพื่อที่นักการศึกษาสามารถวางแผนกำหนดเป้าหมายหรือช่วยเหลือผู้เรียนได้เป็นรายบุคคล

บรรณานุกรม

บรรณานุกรม

ภาษาไทย

กระทรวงศึกษาธิการ. (2553). แนวการจัดการกิจกรรมแนะแนวตามหลักสูตรแกนกลางการศึกษา
ขั้นพื้นฐาน พุทธศักราช 2551. กรุงเทพฯ: โรงพิมพ์ชุมนุมสหกรณ์การเกษตรแห่งประเทศไทย.

กระทรวงศึกษาธิการ. (2559). มาตรฐานการแนะแนว. กรุงเทพฯ: โรงพิมพ์ชุมนุมสหกรณ์การเกษตรแห่งประเทศไทย.

กรมวิชาการ. (2549 ก.) การกำหนดวิสัยทัศน์และยุทธศาสตร์ของการแนะแนวมิติใหม่. กรุงเทพฯ:
โรงพิมพ์คุรุสภาลาดพร้าว.

โกศล มีคุณ. (2544). การแนะแนวจริยธรรมและจรรยาวิชาชีพในประมวลสาระชุดวิชาหลักการและ

แนวคิดทางการแนะแนว หน่วยที่ 1, (น. 1-55). นนทบุรี: มหาวิทยาลัยสุโขทัยธรรมา-

ธิราช สาขาวิชาศึกษาศาสตร์.

โกเมศ อัมพวัน. (ม.ป.ป.) การทำเหมืองข้อมูล. ชลบุรี: สาขาวิทยาการคอมพิวเตอร์ คณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา.

พนม ลิ้มอารีย์. (2548). *การแนะแนวเบื้องต้น*. (พิมพ์ครั้งที่ 2). กรุงเทพฯ: โอเดียนสโตร์.

นงลักษณ์ ประเสริฐ, และจรินทร วินทะวิไชย์. (2548). หลักการแนะแนว. กรุงเทพฯ: คณะครุศาสตร์จุฬาลงกรณ์มหาวิทยาลัย.

นนท์ บุญนิธิประเสริฐ, และ ชัยพร เขมะภาตะพันธ์. (2552). การกรองข้อความภาษาไทยและภาษาอังกฤษของบริการส่งข้อความสั้นบนเครือข่ายโทรศัพท์เคลื่อนที่, *Proceeding of The 7th National Conference on Computing and Information Technology (NCCIT 2009),* (น. 34-39). Bangkok.

ราชบัณฑิตยสถาน. (2556). พจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ. 2554. กรุงเทพฯ: นานมีบุ๊คส์พับลิเคชั่นส์.

วุฒิชัย กำจรกิตติคุณ, และชัยพร เขมะภาตะพันธ์. (2559). การวิเคราะห์ความผิดปกติข้อมูลการใช้ไฟฟ้าด้วยนาอีฟเบย์. *การประชุมวิชาการระดับประเทศด้านเทคโนโลยีสารสนเทศ (NCIT) ครั้งที่ 8* (น. 403-408).

สมาคมโปรแกรมเมอร์ไทย. (2561). อะไรคือการเรียนรู้ของเครื่อง Machine Learning? (ฉบับ

มือใหม่). เข้าถึงได้จาก https://www.thaiprogrammer.org/2018/12/อะไรคือ-การเรียนรู้/

บรรณานุกรม (ต่อ)

ภาษาไทย

สุพัฒน์กุล ภัคโชค. (2556). *ตัวแบบการเลือกแผนการเรียนระดับมัธยมศึกษาตอนปลายโดยการ*

พิจารณาผลการเรียนรายวิชาหลักด้วยเทคนิคเหมืองข้อมูล (วิทยานิพนธ์ปริญญาวิทยา

ศาสตรมหาบัณฑิต). กรุงเทพฯ: มหาวิทยาลัยธุรกิจบัณฑิตย์.

สำนักวิชาการและมาตรฐานการศึกษา. (2554). แผนยุทธศาสตร์และการแนะแนวระดับการศึกษา

ขั้นพื้นฐานในช่วงแผนพัฒนาเศรษฐกิจและสังคมแห่งชาติ. ฉบับที่ 11 (พ.ศ. 2555-

2559). กรุงเทพฯ: สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน กระทรวงศึกษาธิการ.

อตินติ วัฒนบุรานนท์. (2560). *การจำแนกคุณภาพการใช้งานและปัญหาของเกมแอ็คชัน จากบท*

วิจารณ์ของผู้ใช้ด้วยการทำเหมืองข้อความ (วิทยานิพนธ์ปริญญาวิทยาศาสตร

มหาบัณฑิต). กรุงเทพฯ: จุฬาลงกรณ์มหาวิทยาลัย.

อัชรา เอิบสุขสิริ. (2556). จิตวิทยาสำหรับครู. กรุงเทพฯ: สำนักพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

อรทิพย์ เลื่อยงาม, และชัยพร เขมะภาตะพันธ์. (2554). การจัดประเภทเอกสารด้วยวิธีเอสวีเอ็มเพื่อการป้องกันเอกสารรั่วไหล, Proceedings การประชุมวิชาการ “นเรศวรวิจัย” ครั้งที่7 (น. 3-12). กลุ่มวิทยาศาสตร์และเทคโนโลยี.

ภาษาต่างประเทศ

Adcock, J.C. Allen, E. Day, M. Frick, S. Hinchliff, J. Johnson, M. Morley-Short, S. Pallister, S. Price, A.B. and Stanisic, S. (2015). Advances in quantum machine learning. Retrieved from https://www.researchgate.net/publication/286513346

Anaconda Inc. (2021). Created Anaconda & Conda. Retrieved from https://www.anaconda.com/about-us

Drake, Guido Van Rossum Fred L. (2003). An Introduction to Python. Network Theory Limited.

Géron, A. (2017). Hands-on machine learning with Scikit-Learn and TensorFlow. CA. O'Reilly Media.

Géron, A. (2019). Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow. 2nd

edition. CA: O'reilly Media, Inc.

บรรณานุกรม (ต่อ)

ภาษาต่างประเทศ

Han, J. and Kamber, M. (2000). Data Mining Concepts and Techniques. The Morgan Kaufmann

Publishers.

Hunter, J.D. (2008). History of Matplotlib. Retrieved from https://matplotlib.org/stable/users/-

history

Jaroenpuntaruk, V. Wichadakul, D. (2015). Utilizing Data Mining Techniques to Forecast

Student Academic Achievement of Kasetsart University Laboratory School

Kamphaeng Saen Campus Educational Research and Development Center.

Veridian E-J Sci Technol Silpakorn Univ, 2(2), p.p. 1–17.

Joyce, J. (2003). Bayes' Theorem. Retrieved from https://plato.stanford.edu/archives/spr2019/

entries/bayes-theorem/

Miller, C. H. (1976). Foundation of Guidance. New York: Harper&Row Publisher.

Nilsson, N, J. (2005). Introduction to MACHINE LEARNING. CA: Stanford University.

Numpy. (2019). Some Information About the Numpy Project and Community. Retrieved from

https://numpy.org/about/

Pedregosa, F. Varoquaux, G. Gramfort, A. Michel, V. Thirion, B. Grisel, O. Blondel, M.

Prettenhofer, P. Weiss, R. Dubourg, V. Vanderplas, J. Passos, A. Cournapeau, D.

Brucher, M. Perrot, M. and Duchesnay, É. (2011). Scikit-learn: Machine Learning in

Python. Journal of Machine Learning Research. 12(85), p.p. 2825-2830.

Pandas. (n. d.) About Pandas. Retrived from https://pandas.pydata.org/about/Pandas

Perez, F. (2014). Project Jupyter. Retrieved from https://speakerdeck.com/fperez/project-jupyter

Pupale, R. (2018). Support Vector Machines (SVM) – An Overview. Retrieved from

https: / / towardsdatascience. com/ https-medium-com-pupalerushikesh-svm-

f4b42800e989

Python Thailand. from https://python-thailand.github.io/getting-python/library-package.html

Raschka, S. & Mirjalili, V. (2017). Python Machine Learning. 2nd edition. Birmingham:Packt

Publishing Ltd.

บรรณานุกรม (ต่อ)

ภาษาต่างประเทศ

Scikit-Learn Developers. (2020). About us. Retrieved from https://scikit learn.org/stable/-

about.html

Scikit-Learn Developers. (2020). Decision Trees. Retrieved from https://scikit-

learn.org/stable/modules/tree.html#classification

Scikit-Learn Developers. (2020). Decomposing signals in components (matrix factorization

problems. Retrieved from https://scikit-learn.org/stable/modules/

decomposition.html#saprsepca

Scikit-Learn Developers. (2020). Feature selection. Retrieved from https://scikit-learn.org/

stable/modules/feature_selection.html

Walid Mohamed Aly, Osama Fathy Hegazy and Heba Mohmmed Nagy Rashad. (2013).

Automated Student Advisory using Machine Learning. *International Journal of*

Computer Applications, 81(19), p.p. 19-24.

Witten, I. H. and Frank, E. (2005). Data Mining: Practical Machine Learning Tools and

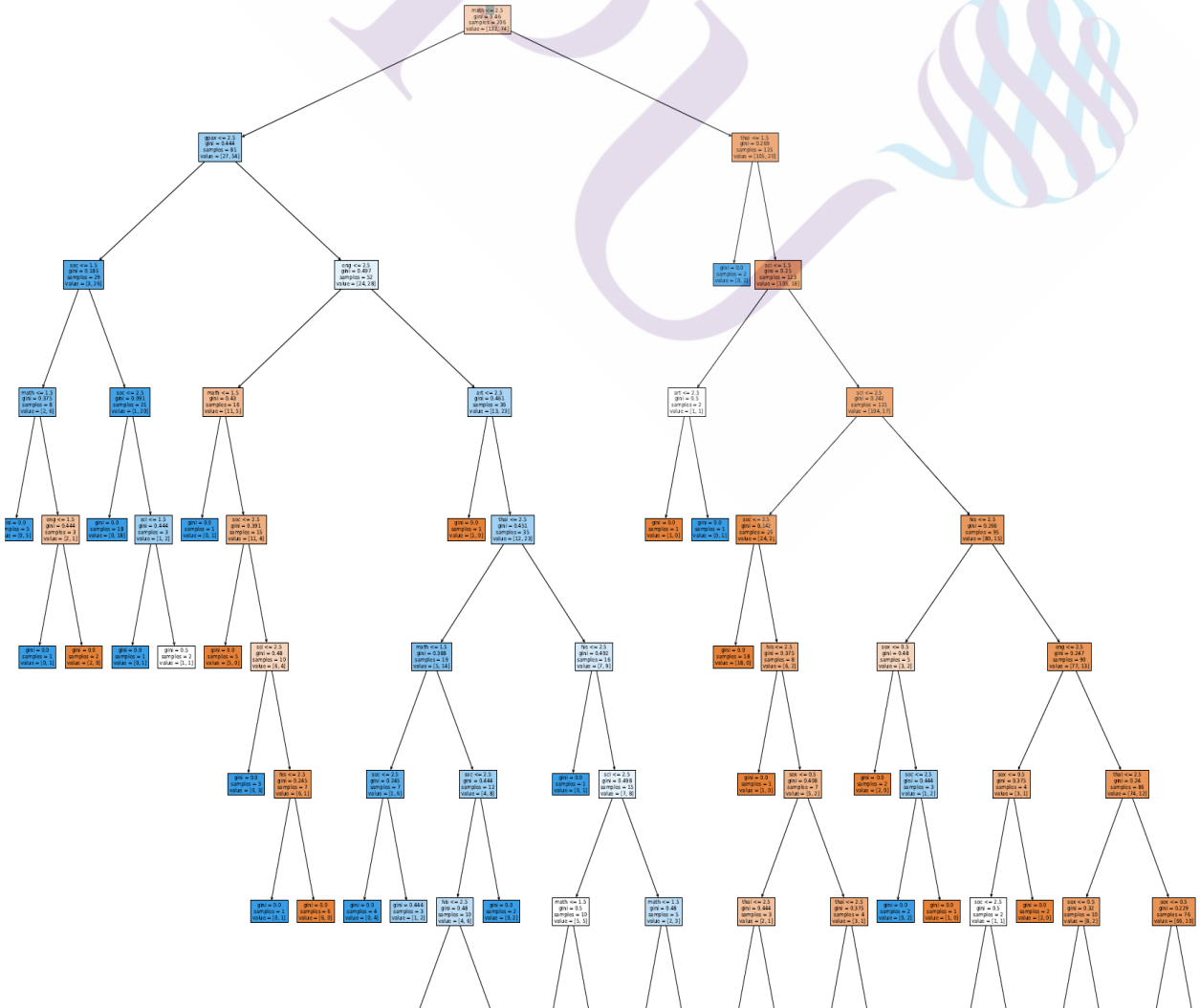
Techniques. The Morgan Kaufmann Publishers.

ภาคผนวก

DPU

ภาคผนวก ก

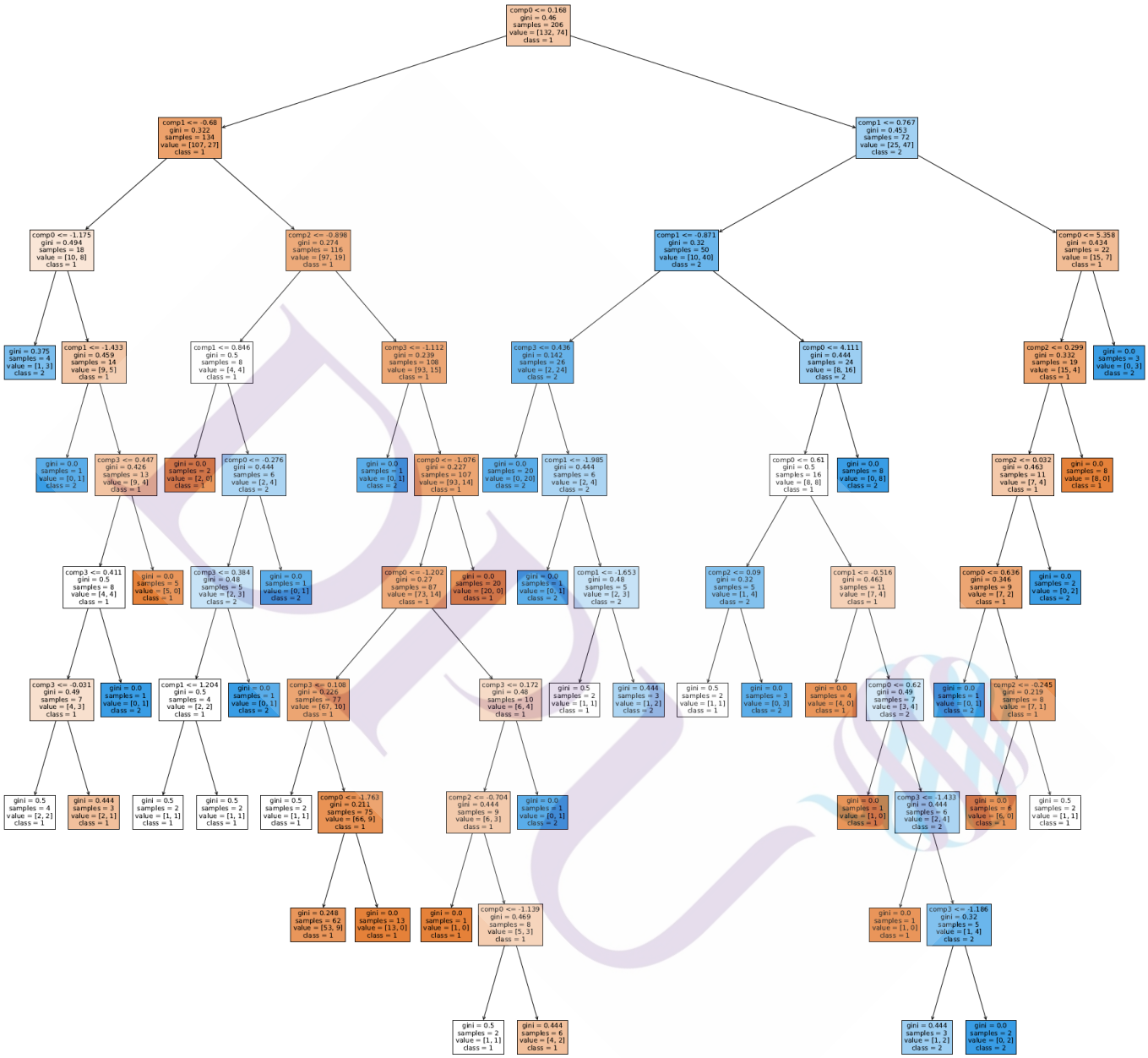
ภาพต้นไม้ตัดสินใจด้วยวิธีการ Correlation Analysis



ภาคผนวก ข

ภาพต้นไม้ตัดสินใจด้วยวิธีการ Principal Component Analysis: PCA

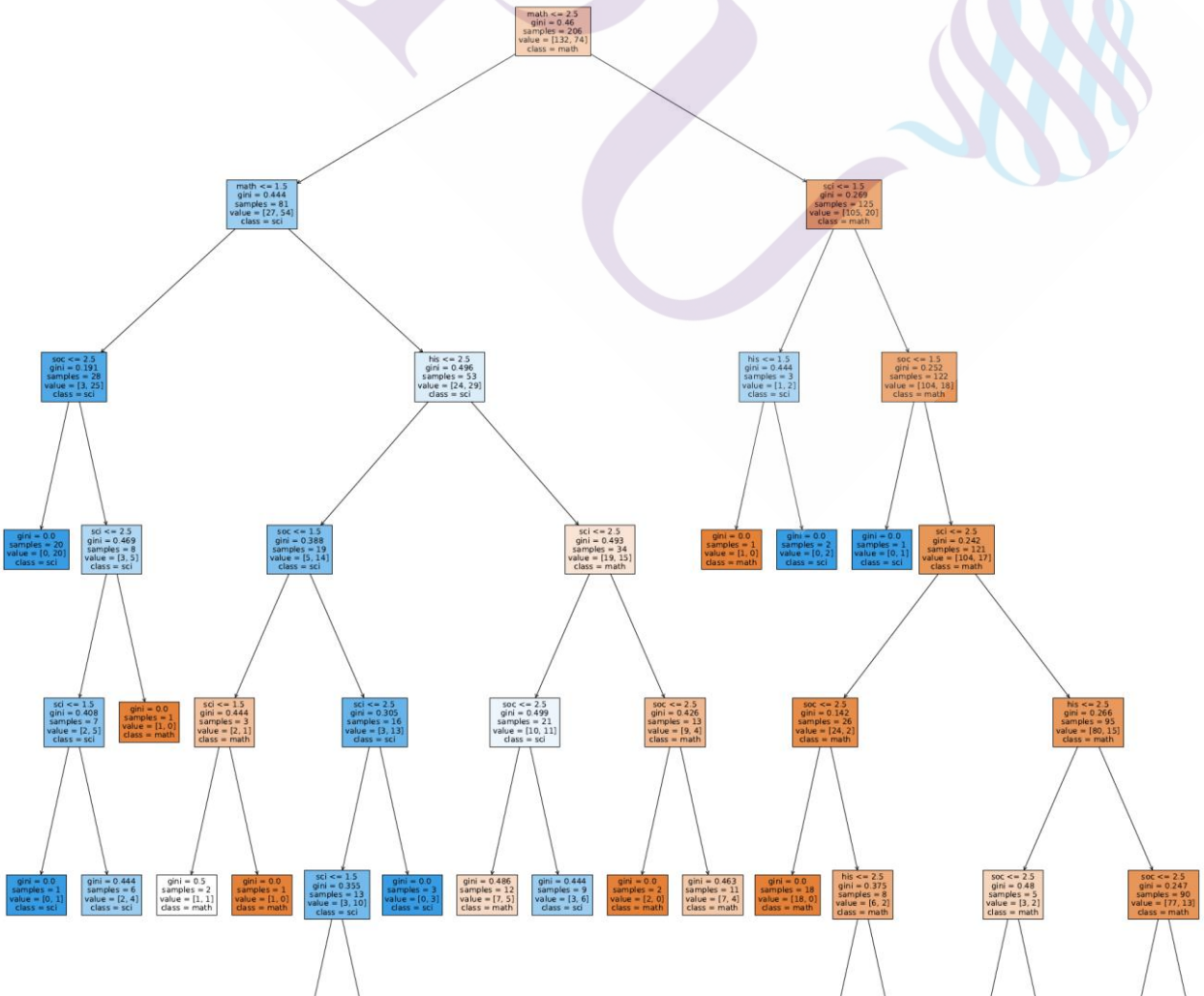
DPU


comp0 <= 0.168
gini = 0.46
samples = 206
value = [132, 74]
class = 1
comp1 <= -0.68
gini = 0.322
samples = 134
value = [107, 27]
class = 1
comp1 <= 0.767
gini = 0.453
samples = 72
value = [25, 47]
class = 2
comp0 <= -1.175
gini = 0.494
samples = 18
value = [10, 8]
class = 1
comp2 <= -0.898
gini = 0.274
samples = 116
value = [97, 19]
class = 1
comp1 <= -0.871
gini = 0.32
samples = 50
value = [10, 40]
class = 2
comp0 <= 5.358
gini = 0.434
samples = 22
value = [15, 7]
class = 1
gini = 0.375
samples = 4
value = [1, 3]
class = 2
comp1 <= -1.433
gini = 0.459
samples = 14
value = [9, 5]
class = 1
comp1 <= 0.846
gini = 0.5
samples = 8
value = [4, 4]
class = 1
comp3 <= -1.112
gini = 0.239
samples = 108
value = [93, 15]
class = 1
comp3 <= 0.436
gini = 0.142
samples = 26
value = [2, 24]
class = 2
comp0 <= 4.111
gini = 0.444
samples = 24
value = [8, 16]
class = 2
comp2 <= 0.299
gini = 0.332
samples = 19
value = [15, 4]
class = 1
gini = 0.0
samples = 3
value = [0, 3]
class = 2
gini = 0.0
samples = 1
value = [0, 1]
class = 2
comp3 <= 0.447
gini = 0.426
samples = 13
value = [9, 4]
class = 1
gini = 0.0
samples = 2
value = [2, 0]
class = 1
comp0 <= -0.276
gini = 0.444
samples = 6
value = [2, 4]
class = 2
gini = 0.0
samples = 1
value = [0, 1]
class = 2
comp0 <= -1.076
gini = 0.227
samples = 107
value = [93, 14]
class = 1
gini = 0.0
samples = 20
value = [0, 20]
class = 2
comp1 <= -1.985
gini = 0.444
samples = 6
value = [2, 4]
class = 2
comp0 <= 0.61
gini = 0.5
samples = 16
value = [8, 8]
class = 1
gini = 0.0
samples = 8
value = [0, 8]
class = 2
comp2 <= 0.032
gini = 0.463
samples = 11
value = [7, 4]
class = 1
gini = 0.0
samples = 8
value = [8, 0]
class = 1
comp3 <= 0.411
gini = 0.5
samples = 8
value = [4, 4]
class = 1
gini = 0.0
samples = 5
value = [5, 0]
class = 1
comp3 <= 0.384
gini = 0.48
samples = 5
value = [2, 3]
class = 2
gini = 0.0
samples = 1
value = [0, 1]
class = 2
comp0 <= -1.202
gini = 0.27
samples = 87
value = [73, 14]
class = 1
gini = 0.0
samples = 20
value = [20, 0]
class = 1
gini = 0.0
samples = 1
value = [0, 1]
class = 2
comp1 <= -1.653
gini = 0.48
samples = 5
value = [2, 3]
class = 2
comp2 <= 0.09
gini = 0.32
samples = 5
value = [1, 4]
class = 2
comp1 <= -0.516
gini = 0.463
samples = 11
value = [7, 4]
class = 1
comp0 <= 0.636
gini = 0.346
samples = 9
value = [7, 2]
class = 1
gini = 0.0
samples = 2
value = [0, 2]
class = 2
comp3 <= -0.031
gini = 0.49
samples = 7
value = [4, 3]
class = 1
gini = 0.0
samples = 1
value = [0, 1]
class = 2
comp1 <= 1.204
gini = 0.5
samples = 4
value = [2, 2]
class = 1
gini = 0.0
samples = 1
value = [0, 1]
class = 2
comp3 <= 0.108
gini = 0.226
samples = 77
value = [67, 10]
class = 1
comp3 <= 0.172
gini = 0.48
samples = 10
value = [6, 4]
class = 1
gini = 0.5
samples = 2
value = [1, 1]
class = 1
gini = 0.444
samples = 3
value = [1, 2]
class = 2
gini = 0.5
samples = 2
value = [1, 1]
class = 1
gini = 0.0
samples = 3
value = [0, 3]
class = 2
gini = 0.0
samples = 4
value = [4, 0]
class = 1
comp0 <= 0.62
gini = 0.49
samples = 7
value = [3, 4]
class = 2
gini = 0.0
samples = 1
value = [0, 1]
class = 2
comp2 <= -0.245
gini = 0.219
samples = 8
value = [7, 1]
class = 1
gini = 0.5
samples = 4
value = [2, 2]
class = 1
gini = 0.444
samples = 3
value = [2, 1]
class = 1
gini = 0.5
samples = 2
value = [1, 1]
class = 1
gini = 0.5
samples = 2
value = [1, 1]
class = 1
gini = 0.5
samples = 2
value = [1, 1]
class = 1
comp0 <= -1.763
gini = 0.211
samples = 75
value = [66, 9]
class = 1
comp2 <= -0.704
gini = 0.444
samples = 9
value = [6, 3]
class = 1
gini = 0.0
samples = 1
value = [0, 1]
class = 2
gini = 0.0
samples = 1
value = [1, 0]
class = 1
comp3 <= -1.433
gini = 0.444
samples = 6
value = [2, 4]
class = 2
gini = 0.0
samples = 4
value = [0, 4]
class = 2
gini = 0.5
samples = 2
value = [1, 1]
class = 1
gini = 0.248
samples = 62
value = [53, 9]
class = 1
gini = 0.0
samples = 13
value = [13, 0]
class = 1
gini = 0.0
samples = 1
value = [1, 0]
class = 1
comp0 <= -1.139
gini = 0.469
samples = 8
value = [5, 3]
class = 1
gini = 0.0
samples = 1
value = [1, 0]
class = 1
comp3 <= -1.186
gini = 0.32
samples = 5
value = [1, 4]
class = 2
gini = 0.5
samples = 2
value = [1, 1]
class = 1
gini = 0.444
samples = 6
value = [4, 2]
class = 1
gini = 0.444
samples = 3
value = [1, 2]
class = 2
gini = 0.0
samples = 2
value = [0, 2]
class = 2

DPU

ภาคผนวก ค

ภาพต้นไม้ตัดสินใจด้วยวิธีการ Feature Selection: Backward selection


math <= 2.5
gini = 0.46
samples = 206
value = [132, 74]
class = math
math <= 1.5
gini = 0.444
samples = 81
value = [27, 54]
class = sci
sci <= 1.5
gini = 0.269
samples = 125
value = [105, 20]
class = math
soc <= 2.5
gini = 0.191
samples = 28
value = [3, 25]
class = sci
his <= 2.5
gini = 0.496
samples = 53
value = [24, 29]
class = sci
his <= 1.5
gini = 0.444
samples = 3
value = [1, 2]
class = sci
soc <= 1.5
gini = 0.252
samples = 122
value = [104, 18]
class = math
gini = 0.0
samples = 20
value = [0, 20]
class = sci
sci <= 2.5
gini = 0.469
samples = 8
value = [3, 5]
class = sci
soc <= 1.5
gini = 0.388
samples = 19
value = [5, 14]
class = sci
sci <= 2.5
gini = 0.493
samples = 34
value = [19, 15]
class = math
gini = 0.0
samples = 1
value = [1, 0]
class = math
gini = 0.0
samples = 2
value = [0, 2]
class = sci
gini = 0.0
samples = 1
value = [0, 1]
class = sci
sci <= 2.5
gini = 0.242
samples = 121
value = [104, 17]
class = math
sci <= 1.5
gini = 0.408
samples = 7
value = [2, 5]
class = sci
gini = 0.0
samples = 1
value = [1, 0]
class = math
sci <= 1.5
gini = 0.444
samples = 3
value = [2, 1]
class = math
sci <= 2.5
gini = 0.305
samples = 16
value = [3, 13]
class = sci
soc <= 2.5
gini = 0.499
samples = 21
value = [10, 11]
class = sci
soc <= 2.5
gini = 0.426
samples = 13
value = [9, 4]
class = math
soc <= 2.5
gini = 0.142
samples = 26
value = [24, 2]
class = math
his <= 2.5
gini = 0.266
samples = 95
value = [80, 15]
class = math
gini = 0.0
samples = 1
value = [0, 1]
class = sci
gini = 0.444
samples = 6
value = [2, 4]
class = sci
gini = 0.5
samples = 2
value = [1, 1]
class = math
gini = 0.0
samples = 1
value = [1, 0]
class = math
sci <= 1.5
gini = 0.355
samples = 13
value = [3, 10]
class = sci
gini = 0.0
samples = 3
value = [0, 3]
class = sci
gini = 0.486
samples = 12
value = [7, 5]
class = math
gini = 0.444
samples = 9
value = [3, 6]
class = sci
gini = 0.0
samples = 2
value = [2, 0]
class = math
gini = 0.463
samples = 11
value = [7, 4]
class = math
gini = 0.0
samples = 18
value = [18, 0]
class = math
his <= 2.5
gini = 0.375
samples = 8
value = [6, 2]
class = math
soc <= 2.5
gini = 0.48
samples = 5
value = [3, 2]
class = math
soc <= 2.5
gini = 0.247
samples = 90
value = [77, 13]
class = math

ประวัติผู้เขียน

ชื่อ – นามสกุล	นายฐิติชัย รักบำรุง
ประวัติการศึกษา	พ.ศ. 2551 ศิลปศาสตรบัณฑิต สาขาสารสนเทศศาสตร์ มหาวิทยาลัยสุโขทัยธรรมาธิราช พ.ศ. 2564 วิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์และโทรคมนาคม มหาวิทยาลัยธุรกิจบัณฑิตย์
ตำแหน่งและสถานที่ทำงานปัจจุบัน	อาจารย์ประจำภาควิชานวัตกรรมและเทคโนโลยีการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยบูรพา

DPU