การศึกษาเทคนิคของต้นไม้ตัดสินใจเพื่อช่วยตัดสินใจในการเลือกสาขาของนักเรียน

STUDY OF THE DECISION TREE TO HELP STUDENT IN DECIDING THE FIELD OF STUDY

โดย

ตะวัน ระวิงทอง

TAWAN RAVINGTHONG

อาจารย์ที่ปรึกษา

รศ.ดร.อาริต ธรรมโน

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน
หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ภาคเรียนที่ 2 ปีการศึกษา 2552

STUDY OF THE DECISION TREE TO HELP STUDENT IN DECIDING THE FIELD OF STUDY

TAWAN RAVINGTHONG

A REPORT SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS OF THE COURSE
SYSTEM DEVELOPMENT PROJECT
MASTER OF SCIENCE PROGRAM IN INFORMATION TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG
2/ 2009

ใบรับรองโครงการพัฒนาระบบงาน

(SYSTEM DEVELOPMENT PROJECT)

เรื่อง

การศึกษาเทคนิคของต้นไม้ตัดสินใจเพื่อช่วยตัดสินใจในการเลือกสาขาของนักเรียน

STUDY OF THE DECISION TREE TO HELP STUDENT IN DECIDING THE FIELD OF STUDY

นางสาวตะวัน ระวิงทอง

รหัสประจำตัว 50066411

ขอรับรองว่ารายงานฉบับนี้ข้าพเจ้าไม่ได้คัดลอกมาจากที่ใด

รายงานฉบับนี้ได้รับการตรวจสอบและอนุมัติให้เป็นส่วนหนึ่งของ

การศึกษาวิชาโครงการพัฒนาระบบงาน หลักสูตรวิทยาศาสตรมหาบัณฑิต (เทคโนโลยีสารสนเทศ)

ภาคเรียนที่ 2 ปีการศึกษา 2552

.....**อาจารย์ที่ปรึกษา**

(รศ.ดร. อาริต ธรรมโน)

.....**กรรมการสอบ**

(รศ.ดร. วรพจน์ กรีสุระเดช)

.....**กรรมการสอบ**

(ผศ.ดร. ภัทรชัย ลลิตโรจน์วงศ์)

หัวข้อ	การศึกษาเทคนิคของต้นไม้ตัดสินใจเพื่อช่วยตัดสินใจในการเลือกสาขาของนักเรียน
นักศึกษา	นางสาวตะวัน ระวิงทอง
รหัสนักศึกษา	50066411
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2552
อาจารย์ที่ปรึกษา	รศ.ดร. อาริต ธรรมโน

บทคัดย่อ

การศึกษาในปัจจุบันมีความสำคัญ และมีการแข่งขันที่สูง จึงมีความจำเป็นที่จะต้องประเมินคุณภาพของนักเรียนแต่ละคนว่าเหมาะที่จะศึกษาในสาขาการเรียนที่ได้ทำการเลือกเพื่อศึกษาต่อในโรงเรียนนั้นๆ หรือไม่ เนื่องจากจะส่งผลโดยตรงกับผู้เรียนเอง เช่น ในกรณีที่ตัดสินใจในการเลือกเรียนสาขาที่ไม่มีความเหมาะสมกับความสามารถของตนเองหรือการมีพื้นฐานของรายวิชาที่มีความจำเป็นในการศึกษาในสาขานั้นๆ อ่อนเกินไป ทำให้เมื่อเข้าไปศึกษาต่อแล้วส่งผลกระทบกับการเรียนของผู้เรียนหรือจบการศึกษาไปอย่างไม่มีคุณภาพ โครงงานนี้จะเป็นการศึกษาเทคนิคของต้นไม้ตัดสินใจเพื่อช่วยตัดสินใจในการเลือกสาขาของนักเรียน ซึ่งจะประกอบด้วยข้อมูลเกี่ยวกับประวัติ และผลการศึกษาของนักเรียนแต่ละคน ซึ่งข้อมูลต่างๆ จะถูกนำไปวิเคราะห์สำหรับช่วยตัดสินใจในการเลือกสาขาการเรียนของนักเรียน เพื่อเป็นประโยชน์แก่ตัวนักเรียน ผู้ปกครอง ในการที่จะเลือกสาขาการเรียนเพื่อเข้าศึกษาต่อ และสำหรับอาจารย์ผู้ที่ดูแลในเรื่องการจัดการแนะแนวด้านการเรียนในการให้ข้อมูลในการศึกษาต่อของนักเรียน และผู้บริหารสถานศึกษา เพื่อนำข้อมูลที่ได้ไปใช้ในการจัดการภายในสถานศึกษาต่อไป

Title	Study of the Decision Tree to Help Student in Deciding the Field of Study
Student	Miss Tawan Ravingthong
Student ID	50066411
Degree	Master of Science
Program	Information Technology
Major	Information Science
Academic Year	2009
Advisor	Assoc. Prof. Dr. Arit Thammano

ABSTRACT

The current study is important and highly competitive. To assess the quality of students in each of the selected majors have quality or not. Because that will directly impact on learners themselves. In such, the decision to choose the major that is not fit to own or have the foundation of the courses are required to study in major that are too weak to make study impact on learning or graduate without quality. This project study of the decision tree to help student in deciding the field of study. Contains information about the history and education of each student. That information will be analysis for decision support in selected major of the student. To benefit for students and parents in selected major for further study. To benefits for teachers in guiding students with education and executives to bring management education within the school in the future.

กิตติกรรมประกาศ

การจัดทำโครงงานพัฒนาระบบนี้สำเร็จลุล่วงได้อย่างดี เพราะได้รับความช่วยเหลือทั้งทางด้านความรู้ แนวทางปฏิบัติ จาก รศ.ดร.อาริต ธรรมโน อาจารย์ที่ปรึกษาโครงการที่ได้ให้คำแนะนำ และข้อคิดเห็นต่างๆ ที่เป็นประโยชน์ต่อแนวทางในการพัฒนาโครงงานมาโดยตลอดจนทำให้โครงการนี้สำเร็จลุล่วงไปได้ด้วยดี ขอขอบพระคุณอาจารย์เป็นอย่างสูง

ขอกราบขอบพระคุณ บิดา มารดา และครอบครัวของข้าพเจ้าที่เป็นกำลังใจ และให้การสนับสนุนในทุกเรื่องๆ ทำให้ข้าพเจ้าสามารถทำโครงงานพัฒนาระบบสำเร็จลุล่วงด้วยดี

ขอบคุณเพื่อนๆ ทุกคนที่คอยให้กำลังใจเสมอมา

ท้ายนี้ขอขอบคุณ สถาบัน คณะ และคณาจารย์ทุกท่านที่ได้ให้ความกรุณาประสิทธิประสาทวิชาความรู้ จนสามารถพัฒนาโครงการพัฒนาระบบจนสำเร็จ

ตะวัน ระวิงทอง

สารบัญ

หน้า

บทคัดย่อภาษาไทย.....I

บทคัดย่อภาษาอังกฤษ.....II

กิตติกรรมประกาศ.....III

สารบัญ.....IV

สารบัญตาราง.....VI

สารบัญรูป.....VII

บทที่ 1 บทนำ.....1

1.1 ความเป็นมาและความสำคัญของปัญหา.....1

1.2 วัตถุประสงค์ของการศึกษา.....2

1.3 ขอบเขตของการศึกษา.....2

1.4 ประโยชน์ที่คาดว่าจะได้รับ.....2

1.5 ขั้นตอนและวิธีการดำเนินงาน.....2

บทที่ 2 ดาต้าไมนิ่ง (Data Mining).....4

2.1 นิยามของดาต้าไมนิ่ง.....4

2.2 กระบวนการในการทำดาต้าไมนิ่ง.....4

2.2.1 การกำหนดวัตถุประสงค์ของงาน.....4

2.2.2 การเตรียมข้อมูล.....4

2.2.3 การทำดาต้าไมนิ่ง.....5

2.2.4 การวิเคราะห์ผลลัพธ์.....6

2.2.5 การนำสารสนเทศไปใช้.....7

2.3 การจำแนกประเภทข้อมูล.....7

2.4 เทคนิคการจำแนกประเภทข้อมูลด้วยแผนภูมิต้นไม้.....8

2.4.1 เทคนิคการสร้างต้นไม้ตัดสินใจ.....8

2.4.2 อัลกอริทึม ID3.....9

2.4.2.1 สมการของอัลกอริทึม ID3.....10

2.4.3 อัลกอริทึม C4.5.....14

2.4.3.1 สมการของอัลกอริทึม C4.5.....15

สารบัญ (ต่อ)

หน้า

2.4.3.2 Over-Fitting 16
2.4.3.3 กรณีที่ข้อมูลเป็นตัวเลขต่อเนื่อง.....17
2.4.3.4 กรณีที่ข้อมูลขาดหายไป 18
2.4.3.5 Tree-Pruning 19

บทที่ 3 การวิเคราะห์และออกแบบ.....20
3.1 การวิเคราะห์และออกแบบระบบ 20
3.1.1 Use-Case Diagram 20
3.1.2 Use-Case Description.....21
3.1.3 Activity Diagram 26
3.1.4 Class Diagram 29
3.1.5 ER Diagram.....30
3.1.6 พจนานุกรมข้อมูล.....31

บทที่ 4 การประยุกต์ใช้ดาต้าไมนิ่งโดยใช้เทคนิคต้นไม้ตัดสินใจ.....34
4.1 การกำหนดวัตถุประสงค์ของงาน.....34
4.2 การเตรียมข้อมูล.....34
4.3 การทำดาต้าไมนิ่ง.....37
4.4 การพัฒนาระบบงาน.....37
4.5 การวิเคราะห์ผลลัพธ์.....44

บทที่ 5 สรุปผลการศึกษาและข้อเสนอแนะ.....46
5.1 สรุปการพัฒนาระบบงาน.....46
5.2 ข้อเสนอแนะ.....46

บรรณานุกรม.....47
ประวัติผู้เขียน.....48

สารบัญตาราง

ตารางที่ หน้า

2.1 แสดงชุดข้อมูลกลุ่มตัวอย่างการตัดสินใจในการเลือกซื้อรถยนต์.....11

2.2 แสดงข้อมูลสำหรับการพิจารณาที่ Attribute “Age” 12

2.3 แสดงข้อมูลสำหรับการพิจารณาที่ Attribute “Income”.....12

2.4 แสดงข้อมูลสำหรับการพิจารณาที่ Attribute “Working” 13

2.5 แสดงข้อมูลสำหรับการพิจารณาที่ Attribute “Credit” 13

3.1 แสดงรายละเอียดการทำงานของ Use-Case Login.....22

3.2 แสดงรายละเอียดการทำงานของ Use-Case Cleaning Data.....23

3.3 แสดงรายละเอียดการทำงานของ Use-Case Create Training Data and Build Model.....24

3.4 แสดงรายละเอียดการทำงานของ Use-Case Test Model 25

3.5 แสดงข้อมูล Entity ของนักเรียน 31

3.6 แสดงข้อมูล Entity ของสาขาการเรียน 31

3.7 แสดงข้อมูล Entity ของคะแนนเฉลี่ยตลอดหลักสูตร.....31

3.8 แสดงข้อมูล Entity ของผลการเรียน.....32

3.9 แสดงข้อมูล Entity ของสิทธิผู้ใช้งาน 32

3.10 แสดงข้อมูล Entity ของโมเดล 32

3.11 แสดงข้อมูล Entity ของทรี (Tree).....33

3.12 แสดงข้อมูล Entity ของกฎ 33

4.1 แสดงรายละเอียดของแต่ละ Entity 34

4.2 แสดงคาต้าดิกชันนารีของ Entity Students 34

4.3 แสดงคาต้าดิกชันนารีของ Entity ResultStudy 35

4.4 แสดงคาต้าดิกชันนารีของ Entity Major.....36

4.5 แสดงการกำหนดรหัสสาขาการเรียน.....36

4.6 แสดงการแบ่งคลาสของคะแนนผลการเรียนเฉลี่ย.....36

4.7 แสดงการแบ่งคลาสของผลการเรียนเฉลี่ยแต่ละกลุ่มวิชา.....36

4.8 แสดงตัวอย่างข้อมูลนักเรียน.....37

สารบัญรูป

รูปที่ หน้า

2.1 แสดงขั้นตอนการสร้างโมเดลของการจำแนกประเภทข้อมูล.....8

2.2 โครงสร้างของ Decision Tree 9

2.3 แสดงต้นไม้ตัดสินใจในการซื้อรถยนต์.....14

2.4 แสดงการแบ่งกลุ่มข้อมูลเริ่มต้นที่ Attribute “Age” 16

2.5 ความถูกต้องจากการแยกแยะข้อมูลของต้นไม้ตัดสินใจเทียบกับขนาดของต้นไม้ตัดสินใจ....17

3.1 แสดง Use Case ของระบบ.....20

3.2 แสดง Activity Diagram ของ Login.....26

3.3 แสดง Activity Diagram ของ Cleaning Data.....27

3.4 แสดง Activity Diagram ของ Create Training Data and Build Model.....27

3.5 แสดง Activity Diagram ของ Test Model 28

3.6 แสดง Class Diagram ของระบบการศึกษาเทคนิคของต้นไม้ตัดสินใจเพื่อช่วยตัดสินใจในการเลือกสาขาของนักเรียน.....29

3.7 แสดง ER Diagram ของระบบการศึกษาเทคนิคของต้นไม้ตัดสินใจเพื่อช่วยตัดสินใจในการเลือกสาขาของนักเรียน.....30

4.1 หน้าจอแสดงส่วน Login 38

4.2 หน้าจอแสดง Models 38

4.3 หน้าจอแสดงส่วน Training 39

4.4 หน้าจอแสดงส่วน Missing Values 39

4.5 หน้าจอแสดงการสร้างโมเดล Tree ในกรณีเลือกอัลกอริทึมID3.....40

4.6 หน้าจอแสดงการสร้างโมเดล Tree ในกรณีเลือกอัลกอริทึม C4.5.....40

4.7 หน้าจอแสดงการทดสอบโมเดลที่ได้จากการสร้างโมเดล Tree 41

4.8 หน้าจอแสดงการทดสอบโมเดลที่ได้จากการสร้างโมเดล Tree กรณีทดสอบเป็นรายบุคคล....41

4.9 หน้าจอแสดงข้อมูลกรณีทดสอบเป็นกลุ่ม 42

4.10 หน้าจอแสดงการเลือกข้อมูลที่จะนำมาทดสอบทดสอบ 42

4.11 หน้าจอแสดงการทดสอบโมเดลที่ได้จากการสร้างโมเดล Tree ทดสอบเป็นกลุ่ม 43

4.12 หน้าจอแสดงการทำนายแนวโน้มของผลการเรียนโดยระบุสาขาที่เลือก.....43

4.13 หน้าจอแสดงการทดสอบโมเดลที่ได้จากการสร้างโมเดล Tree.....44

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การศึกษาในปัจจุบันมีความสำคัญ และมีการแข่งขันที่สูง ซึ่งในแต่ละปีจะมีการรับนักเรียนเข้าศึกษาต่อเป็นจำนวนมาก จึงมีความจำเป็นที่จะต้องประเมินคุณภาพของนักเรียนแต่ละคนว่าเหมาะที่จะศึกษาต่อในสาขาการเรียนที่ได้ทำการเลือกในโรงเรียนนั้นๆ หรือไม่ ซึ่งอาจจะส่งผลกระทบกับผู้เรียนโดยตรง โดยเฉพาะปัญหาที่พบมากสำหรับการเลือกสาขาการเรียนของนักเรียนเพื่อเข้าศึกษาต่อในโรงเรียนต่างๆ คือ กรณีที่นักเรียนตัดสินใจเลือกเรียนในสาขาการเรียนที่ไม่มีความเหมาะสมกับความสามารถของตนเองอย่างแท้จริงหรือการมีพื้นฐานของรายวิชาที่มีความจำเป็นในการศึกษาในสาขานั้นๆ อ่อนเกินไป หรือนักเรียนบางคนไม่ทราบว่าตนเองควรจะเลือกเรียนสาขาไหนดีก็อาจจะเลือกเรียนตามเพื่อนๆ ทำให้เมื่อเข้าไปศึกษาตามสาขาที่ได้เลือกแล้วส่งผลกระทบกับการเรียนหรือผลการเรียนตกต่ำ ทำให้นักเรียนไม่สามารถที่จะสำเร็จการศึกษาหรือจบการศึกษาไปอย่างไม่มีคุณภาพ หรืออาจส่งผลกระทบทางด้านจิตใจของผู้เรียนเกิดความไม่อยากเรียน ซึ่งจะส่งผลกระทบกับการศึกษาต่อในระดับสูงต่อไป

สำหรับโครงงานนี้ได้ทำการศึกษาเทคนิคของต้นไม้ตัดสินใจเพื่อช่วยตัดสินใจในการเลือกสาขาของนักเรียน โดยนำเทคนิคของอัลกอริทึม ID3 และอัลกอริทึม C4.5 ซึ่งเป็นอัลกอริทึมหนึ่งในเทคนิคของต้นไม้ตัดสินใจ (Decision Tree) ที่ใช้การจำแนกประเภทข้อมูล (Classification) จากชุดข้อมูลมาช่วยในการวิเคราะห์ข้อมูลสำหรับช่วยตัดสินใจในการเลือกสาขาของนักเรียนแต่ละคนว่าเหมาะสมที่จะเลือกเรียนในสาขาการเรียนไหนมากที่สุด และใช้ในการประเมินความรู้ความสามารถของนักเรียน ซึ่งโดยทั่วไป โรงเรียนต่างๆจะมีการจัดเก็บข้อมูลของนักเรียนไว้เป็นจำนวนมาก เช่น ข้อมูลเกี่ยวกับประวัติ และผลการศึกษาของนักเรียนแต่ละคน แต่ส่วนใหญ่ไม่ได้ถูกนำมาใช้ประโยชน์มากนัก ซึ่งข้อมูลเหล่านั้นสามารถที่จะนำมาประมวลผลเพื่อค้นหาสารสนเทศที่มีประโยชน์ได้ โดยสามารถนำไปวิเคราะห์เพื่อช่วยตัดสินใจในการเลือกสาขาของนักเรียนของโรงเรียนต่างๆ เพื่อเป็นประโยชน์แก่ตัวนักเรียน ผู้ปกครอง ในการที่จะเลือกสาขาการเรียนเพื่อเข้าศึกษาต่อ และสำหรับอาจารย์ผู้ที่ดูแลในเรื่องการจัดการแนะแนวด้านการเรียนในการให้ข้อมูลในการศึกษาต่อของนักเรียน และผู้บริหารสถานศึกษา เพื่อนำข้อมูลที่ได้ไปใช้ในการจัดการภายในสถานศึกษาต่อไป

1.2 วัตถุประสงค์ของการศึกษา

โครงการพัฒนาระบบเพื่อช่วยตัดสินใจในการเลือกสาขาของนักเรียน มีวัตถุประสงค์ ดังนี้

1. เพื่อศึกษาหลักการสร้างต้นไม้ตัดสินใจ (Decision Tree) ด้วยอัลกอริทึม ID3 และอัลกอริทึม C4.5

2. เพื่อนำความรู้ และเทคนิคที่ศึกษามาประยุกต์ใช้กับการวิเคราะห์ข้อมูลผ่านดาต้าไมนิ่ง (Data Mining) เพื่อสร้างแบบจำลองเพื่อช่วยตัดสินใจในการเลือกสาขาของนักเรียน

3. เพื่อพัฒนาโปรแกรมเพื่อช่วยตัดสินใจในการเลือกสาขาของนักเรียน โดยใช้แบบจำลองที่สร้างขึ้น

1.3 ขอบเขตของการศึกษา

1. สร้างแบบจำลองเพื่อช่วยตัดสินใจในการเลือกสาขาของนักเรียน จากเทคนิคการทำเหมืองข้อมูลต้นไม้ตัดสินใจ (Decision Tree) ด้วยอัลกอริทึม ID3 และอัลกอริทึม C4.5 ผ่านการทำงานของโปรแกรมแอพพลิเคชั่น

2. สามารถวิเคราะห์ข้อมูล และแยกประเภทข้อมูลของนักเรียนโรงเรียนพิบูลวิทยาลัย จังหวัดลพบุรี ด้วยแบบจำลองที่สร้างขึ้นตามจุดมุ่งหมายที่ได้กำหนดไว้

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1. เข้าใจหลักการการจำแนกประเภทข้อมูลรวมถึงแนวคิดและทฤษฎีของต้นไม้ตัดสินใจ

2. สามารถสร้างโปรแกรมที่ช่วยสนับสนุนการตัดสินใจในการเลือกสาขาของนักเรียน ซึ่งสามารถนำไปใช้ในการวางแผนการเรียนในอนาคตหรือระดับสูงต่อไป

3. สามารถสร้างโปรแกรมที่ช่วยสนับสนุนการตัดสินใจสำหรับอาจารย์เพื่อแนะแนวทางหรือให้คำปรึกษานักเรียนที่จะเลือกเรียนในสาขาต่างๆ ได้ดีขึ้น และเป็นแนวทางในการจัดการภายในสถานศึกษาต่อไป

4. เพื่อนำความรู้ และเทคนิคที่ศึกษาไปประยุกต์ใช้ในการวิเคราะห์ข้อมูลผ่านดาต้าไมนิ่งโดยการใช้อัลกอริทึม ID3 และอัลกอริทึม C4.5

1.5 ขั้นตอนและวิธีการดำเนินงาน

การพัฒนาโครงการนี้ มีขั้นตอนและวิธีการดำเนินงาน ดังนี้

1. ศึกษาทฤษฎีที่เกี่ยวข้องกับดาต้าไมนิ่ง (Data Mining) การจำแนกประเภท (Classification) เทคนิคของต้นไม้ตัดสินใจ (Decision Tree) และอัลกอริทึม ID3, อัลกอริทึม C4.5 ซึ่งเป็นอัลกอริทึมหนึ่งในเทคนิคของต้นไม้ตัดสินใจ (Decision Tree)

2. การวิเคราะห์และออกแบบโปรแกรม

3. การสร้างและทดสอบโปรแกรม

4. บทสรุปผลการศึกษา และข้อเสนอแนะ

บทที่ 2
ดาต้าไมนิ่ง (Data Mining)

2.1 นิยามของดาต้าไมนิ่ง (Data Mining)

ดาต้าไมนิ่ง (Data Mining) เป็นกระบวนการในการกลั่นกรองข้อมูลจากฐานข้อมูลขนาดใหญ่ (Large Information) โดยใช้การวิเคราะห์ทางสถิติ และเทคนิคแบบจำลองในการหาความสัมพันธ์และรูปแบบทั้งหมดซึ่งมีอยู่จริงในฐานข้อมูล เพื่อให้ได้สารสนเทศที่สามารถนำไปใช้งานได้ (Useful Information) โดยเป็นสารสนเทศที่มีเหตุผล (Valid) เชื่อถือได้ เพื่อนำไปใช้ประโยชน์ในการวิเคราะห์ หรือทำนายสิ่งต่าง ๆ ที่จะเกิดขึ้นโดยอาศัยข้อมูลในอดีต และเป็นข้อมูลที่ช่วยสนับสนุนการตัดสินใจให้กับองค์กรในด้านต่างๆ ต่อไปในอนาคต

2.2 กระบวนการในการทำดาต้าไมนิ่ง (Data Mining)

กระบวนการในการทำดาต้าไมนิ่ง (Data Mining) ประกอบด้วย 5 ขั้นตอน คือ

2.2.1 การกำหนดจุดมุ่งหมายขององค์กรหรือวัตถุประสงค์ของงาน (Business Objective Determination)

เข้าใจปัญหา และสามารถกำหนดความต้องการหรือวัตถุประสงค์ขององค์กรที่ชัดเจนได้ ทำให้สามารถกำหนดวัตถุประสงค์ของการทำ Mining ข้อมูลได้ สามารถเป็นแนวทางในการระบุถึงอัลกอริทึม (Algorithm) และข้อมูลที่สัมพันธ์กับวัตถุประสงค์ขององค์กรได้

2.2.2 การเตรียมข้อมูล (Data Preparation)

กระบวนการในการจัดการข้อมูลสารสนเทศให้อยู่ในรูปแบบมาตรฐาน เพื่อให้สามารถนำเข้าสู่อัลกอริทึมของดาต้าไมนิ่ง (Data Mining) ได้ ซึ่งสามารถแบ่งออกเป็นขั้นตอนย่อยๆ ดังนี้

2.2.2.1 การเลือกข้อมูล (Data Selection)

เป็นการคัดเลือกข้อมูลสำหรับสำหรับการที่จะทำ Mining ในขั้นต่อๆ ไป โดยต้องคำนึงถึงวัตถุประสงค์ของแต่ละองค์กร และลักษณะงานที่จะถูกนำมาใช้ด้วย ซึ่งสามารถแบ่งข้อมูลได้ 2 ลักษณะ คือ

2.2.2.1.1 ข้อมูลที่แบ่งตามปริมาณ (Quantitative Data) หรือ ข้อมูลที่เป็นตัวเลข จะมีความแตกต่างระหว่างค่าที่เป็นไปได้ แบ่งได้ 2 ประเภท คือ

- Discrete คือ ค่าที่เก็บเป็นเลขจำนวนเต็ม เช่น จำนวนคนหรือสิ่งของ เวลา (ปี, เดือน)
- Continuous คือ ค่าที่เก็บเป็นเลขจำนวนจริง เช่น รายได้เฉลี่ย ค่าเฉลี่ยต่างๆ

2.2.2.1.2 ข้อมูลที่แบ่งตามประเภทหรือแบ่งเป็นกลุ่ม (Catagorical) แบ่งได้ 2 ประเภท คือ

- ข้อมูลที่ไม่มีลำดับ (Nominal) หรือ ลำดับไม่มีความสำคัญ เช่น สถานภาพสมรส (โสด, แต่งงาน, หย่าร้าง) เพศ (ชาย, หญิง) ระดับการศึกษา (ปริญญาตรี, ปริญญาโท)
- ข้อมูลที่มีลำดับ (Ordinal) หรือ ลำดับมีความสำคัญ เช่น เครดิตของลูกค้า (ดี, ปานกลาง, แย่) เกรดวิชา (A, B, C)

2.2.2.2 การกลั่นกรองข้อมูลหรือการตรวจสอบข้อมูล (Data Preprocessing)

เป็นกระบวนการที่ทำให้เกิดความมั่นใจว่าคุณภาพของข้อมูลที่ถูกเลือกนั้นถูกต้องเหมาะสมที่จะนำมาใช้ในการวิเคราะห์สำหรับการทำดาต้าไมนิ่ง (Data Mining) โดยใช้หลักการทางสถิติ เช่น การวัดการกระจายของข้อมูลหรือนำข้อมูลนั้นมาสร้างเป็นกราฟ เพื่อช่วยให้เห็นความผิดปกติของข้อมูล ถ้าข้อมูลอยู่ในลักษณะที่เป็นตัวเลข ก็สามารถวิเคราะห์โดยการหาค่าสูงสุด ค่าต่ำสุด ค่าเฉลี่ย และจัดการกับข้อมูลที่ไม่สมบูรณ์หรือข้อมูลที่ผิดปกติให้อยู่ในค่าที่เหมาะสม

- Noisy Data คือ ข้อมูลที่มีความคลาดเคลื่อน เป็นข้อมูลที่มีลักษณะแตกต่างจากที่คาดการณ์ไว้หรือค่าของข้อมูลผิดไปจากค่าที่ควรจะเป็น เช่น การบันทึกข้อมูลรายได้ติดลบ การบันทึกข้อมูลอายุคนเป็น 300 ปี ซึ่งค่าเหล่านี้ต้องทำการแก้ไขให้ถูกต้องหรือไม่นำค่ามาใช้ในการวิเคราะห์
- Missing Value คือ ข้อมูลที่ขาดหายไปหรือข้อมูลบางส่วนหายไป ถ้าข้อมูลหายไปมีจำนวนน้อย สามารถที่จะตัดข้อมูลนั้นทิ้งไปได้ แต่ถ้าข้อมูลที่ขาดหายไปมีจำนวนมาก อาจแทนค่าที่หายไปด้วยค่าเฉลี่ย (Mean) หรือค่าที่ปรากฏบ่อย (Mode) หรือบันทึกเป็น "UNKNOWN" ในการทำนายค่า

2.2.2.3 การแปลงรูปแบบข้อมูล (Data Transformation) เป็นการแปลงข้อมูล หรือจัดข้อมูลให้อยู่ในรูปแบบข้อมูลที่เหมาะสมกับอัลกอริทึม (Algorithm) ในแต่ละเทคนิคของดาต้าไมนิ่ง (Data Mining) ที่เลือกใช้ เช่น การแทนเพศชาย (Male) ด้วย M แทนเพศหญิง (Female) ด้วย F , แปลงข้อมูลตัวเลขให้เป็นช่วงเพื่อใช้สร้างต้นไม้ตัดสินใจ (Decision Tree) ซึ่งการแปลงข้อมูลนั้นมีหลายแบบ เช่น Generalize, Normalize data และ Discretization

2.2.3 การทำดาต้าไมนิ่ง (Data Mining)

เป็นขั้นตอนการเลือกเทคนิคต่างๆ ในการทำดาต้าไมนิ่ง (Data Mining) หรือ สร้างแบบจำลอง (Model) เพื่อให้ได้รูปแบบของข้อมูลที่สามารถนำมาใช้ในการวิเคราะห์หรือทำนายสิ่งต่างๆ ที่จะเกิดขึ้น และสนับสนุนการตัดสินใจในอนาคต โดยจุดประสงค์ในการทำดาต้าไมนิ่ง (Data Mining) มี 2 ประเภท คือ

- Prediction คือ เป็นการคาดคะเนหรือทำนายค่าของข้อมูลในอนาคตที่จะเกิดขึ้น โดยใช้ข้อมูลที่ผ่านมาในอดีต

- Description คือ การหาแบบจำลอง (Model) หรือ รูปแบบ (Pattern) เพื่ออธิบายลักษณะบางอย่างของข้อมูลในรูปแบบที่สามารถเข้าใจได้ง่าย โดยส่วนใหญ่จะเป็นลักษณะการแบ่งกลุ่มให้กับข้อมูล

การทำงานของดาต้าไมนิ่ง (Data Mining Operations) มีดังนี้

1. Database Segmentation เป็นกระบวนการในการแบ่งกลุ่มข้อมูลออกเป็นกลุ่มย่อยๆ ตามลักษณะที่เหมือนกันหรือคล้ายคลึงกันหรือแตกต่างกัน เพื่อให้ง่ายต่อการวิเคราะห์ข้อมูลหรือประยุกต์ในการใช้งานด้านต่างๆ เช่น การแบ่งกลุ่มสาขาการเรียนของนักเรียนตามอายุ เกรดเฉลี่ย

2. Predictive Modeling เป็นกระบวนการในการสร้างแบบจำลองเพื่อทำนายแนวโน้มข้อมูลในอนาคต จากการวิเคราะห์ข้อมูลในอดีตและปัจจุบัน โดยสามารถสำรวจจุดเด่นของข้อมูลที่ปรากฏออกมา และทำการกำหนดจุดเด่นนั้น ซึ่งเป็นตัวที่ใช้แบ่งกลุ่ม เพื่อทำนายว่าข้อมูลอยู่กลุ่มใด เช่น การจัดกลุ่มผู้ป่วยตามผลของการใช้ยารักษา เพื่อระบุรูปแบบการรักษาให้กับผู้ป่วยใหม่ที่เข้ารับการรักษาต่อไป แบ่งเป็น 2 ประเภท คือ

- Classification เป็นการสร้างแบบจำลองโดยใช้ข้อมูลจากกลุ่มที่ได้ทำการกำหนดไว้ล่วงหน้าแล้วสำหรับจัดกลุ่มของข้อมูลให้กับแต่ละข้อมูลในฐานข้อมูล โดยมีการระบุค่าหรือลักษณะที่เป็นไปได้ของข้อมูลภายในแต่ละกลุ่มว่าควรจะอยู่กลุ่มไหน เหมาะกับการใช้ทำนายข้อมูลที่มีค่าไม่ต่อเนื่องหรือ Nominal Value (ลำดับไม่มีความสำคัญ) เช่น การจัดกลุ่มของนักเรียนว่า (เก่ง, ปานกลาง, ไม่เก่ง) โดยพิจารณาจากประวัติและผลการเรียน เทคนิคที่ใช้ เช่น Tree Induction, Neural Induction

- Value Prediction หรือ Regression ใช้สำหรับทำนายหรือประเมินค่าของข้อมูลที่ต่อเนื่องหรือเรียงลำดับ เช่น การทำนายหุ้น เทคนิคที่ใช้จะเป็นเทคนิคทางด้านสถิติ เช่น Linear regression

3. Link Analysis เป็นการวิเคราะห์หาความสัมพันธ์ของข้อมูลภายในกลุ่มหรือระหว่างกลุ่ม เพื่อใช้ลักษณะของข้อมูลหนึ่งไปหาความสัมพันธ์ของอีกข้อมูลหนึ่ง เช่น การระบุว่าในกลุ่มลูกค้าที่ซื้อนมนั้น จะมีลูกค้ากี่เปอร์เซ็นต์ที่ซื้อขนมปังด้วย

4. Deviation Detection เป็นกระบวนการที่ใช้เทคนิคการแสดงลักษณะหรือค้นหาข้อมูลที่มีความแตกต่างจากกลุ่มอื่นหรือผิดไปจากที่คาดเอาไว้ โดยมีการแสดงผลที่สามารถทำความเข้าใจและแปลความหมายได้ง่าย เช่น กราฟ แผนภูมิ

2.2.4 การวิเคราะห์ผลลัพธ์ (Analysis of Results)

การวิเคราะห์และประเมินผลลัพธ์ที่ได้จากการทำดาต้าไมนิ่ง (Data Mining) ว่ามีความเหมาะสมหรือตรงกับวัตถุประสงค์ที่ต้องการหรือไม่ แล้วสรุปความหมายของผลลัพธ์ที่ได้ ซึ่งจะเป็นข้อมูลความรู้ (Knowledge) นำไปเป็นสารสนเทศที่ช่วยในการตัดสินใจ แต่หากผลลัพธ์ที่ได้มาไม่เป็นไปตามวัตถุประสงค์ที่วางไว้ จะย้อนกลับไปทำขั้นตอนก่อนหน้าเพื่อแก้ไขข้อมูลในขั้นตอนนั้น

2.2.5 การนำสารสนเทศไปใช้ (Assimilation of Knowledge)

เป็นขั้นตอนสุดท้ายของกระบวนการทำดาต้าไมนิ่ง(Data Mining) จะนำผลลัพธ์ที่ได้ซึ่งเป็นสารสนเทศที่มีความถูกต้องไปใช้ตามวัตถุประสงค์ที่วางไว้ เพื่อเป็นประโยชน์ในการตัดสินใจและกำหนดวัตถุประสงค์ในอนาคตขององค์กรต่อไปได้

2.3 การจำแนกประเภทข้อมูล (Classification)

เทคนิคการจำแนกประเภทข้อมูล (Data Classification) เป็นกระบวนการจัดกลุ่มข้อมูลหรือจำแนกกลุ่มข้อมูลด้วยคุณลักษณะต่างๆ ที่ได้มีการกำหนดไว้แล้ว โดยขึ้นอยู่กับลักษณะของวัตถุประสงค์นั้นๆ เพื่อสร้างโมเดล (Model) ออกมาเป็นชุดข้อมูล (Class) ซึ่งลักษณะของ Class จะถูกอธิบายโดยกลุ่มของคุณสมบัติ (Attribute) และกลุ่มของข้อมูล (Training data set) ที่ใช้ในการสร้างโมเดลในการจำแนกประเภทข้อมูล ซึ่งโมเดลที่ได้จากการจำแนกประเภทข้อมูลจะทำให้สามารถพิจารณาคลาสสำหรับข้อมูลที่ยังมิได้แบ่งกลุ่มในอนาคตได้ โดยทั่วไปมีรูปแบบการเขียนกฎการจำแนกดังนี้

$$\underbrace{P_1, P_2, \ldots\ldots\ldots, P_r}_{\text{Attribute condition}} \longrightarrow \underbrace{C_i}_{\text{Class}}$$

P_r = เงื่อนไขของกฎ

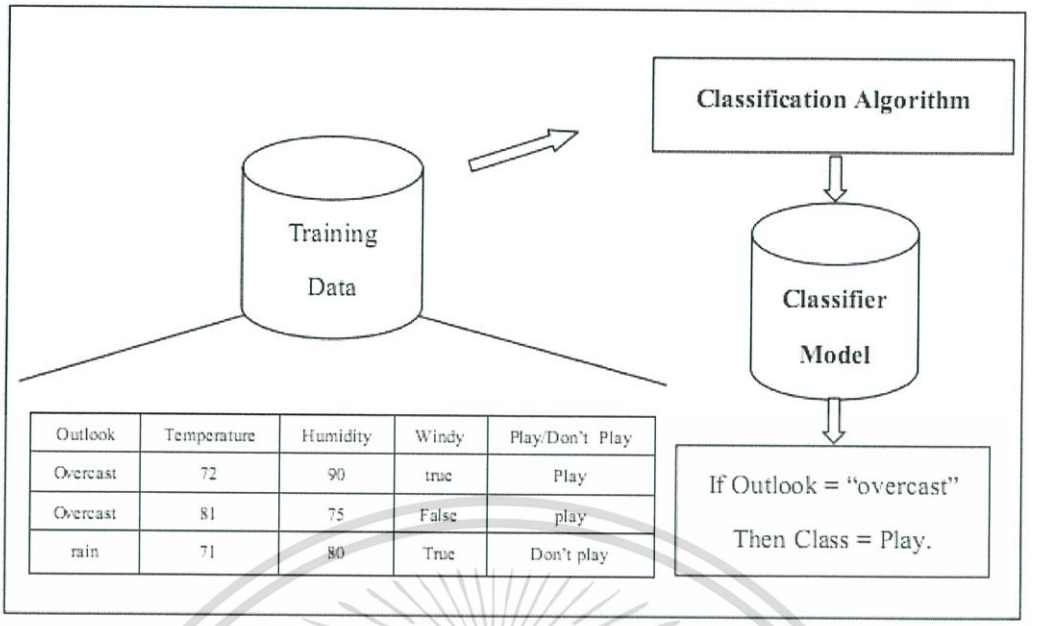
C_i = คลาสของกฎ

หรือ

IF <Condition> THEN <Class> “ถ้า <เงื่อนไข> แล้ว <คลาส>”

ขั้นตอนในการจำแนกประเภทข้อมูลมี 2 ขั้นตอน คือ

1. การสร้างโมเดลต้นแบบ (Classifier Model) เป็นการนำชุดข้อมูล (Training data set) มาผ่านกระบวนการ Classification Algorithm เช่น ID3, C4.5 ซึ่งผลลัพธ์ที่ได้จะอยู่ในรูปของโมเดลของการจำแนกประเภทข้อมูล เช่น ต้นไม้ตัดสินใจ (Tree) ซึ่งสามารถสร้างเป็นกฎได้ ดังรูปที่ 2.1



รูปที่ 2.1 แสดงขั้นตอนการสร้างโมเดลของการจำแนกประเภทข้อมูล (Classifier Model)

2. การใช้โมเดลเพื่อการทำนายแนวโน้มของข้อมูลใหม่ที่จะเกิดขึ้นในอนาคต (Testing data) ซึ่งเป็นข้อมูลที่ยังไม่เคยทำการจัดกลุ่มไว้ โดยการนำข้อมูล (Testing data) มาเปรียบเทียบเทียบกับโมเดลต้นแบบ (Classifier Model) ที่ได้ทำจัดกลุ่มโดยใช้ข้อมูลจากกลุ่มที่ได้ทำการกำหนดไว้ล่วงหน้าแล้ว เพื่อวิเคราะห์ความเป็นไปได้ และเป็นแนวทางในการตัดสินใจในการจัดกลุ่มของข้อมูลนั้นๆ

2.4 เทคนิคการจำแนกประเภทข้อมูลด้วยแผนภูมิต้นไม้

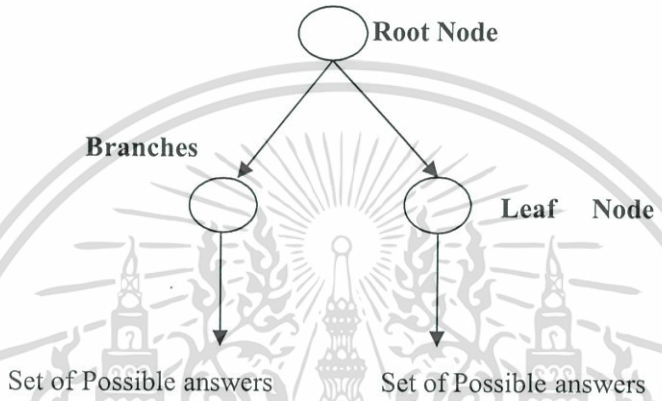
เป็นเทคนิคการนำความรู้ที่เรียนรู้จากข้อมูลที่มีอยู่มาใช้ให้เกิดประโยชน์ และใช้ทำนายแนวโน้มการเกิดขึ้นของข้อมูลที่ยังไม่เกิดขึ้น โดยสามารถสรุปความสัมพันธ์ของข้อมูลออกมาในรูปแบบของกฎหรือแผนภูมิได้

2.4.1 เทคนิคการสร้างต้นไม้ตัดสินใจ (Decision Tree)

Decision Tree เป็นการสร้างต้นไม้การตัดสินใจในการจัดแบ่งกลุ่มจากชุดข้อมูล โดยมีการสร้างการแสดงผลที่เป็น Flow Chart ที่มีโครงสร้างเป็น Tree ซึ่งโครงสร้างจะเป็นลำดับขั้น และจะใช้กฎในการจำแนกประเภทข้อมูล (Classification) คือ "If...Then" โดยโครงสร้างของ Tree ประกอบด้วย

- Decision Node : ส่วนของเงื่อนไขการตัดสินใจ
- Leaf Node : แสดงชื่อ Class ซึ่งเป็นผลลัพธ์ของ Target Attribute จากเงื่อนไขการตัดสินใจ
- Branch : แสดงค่าที่เป็นไปได้ของแต่ละ Attribute โดยการเชื่อมต่อระหว่าง Node

การสร้างต้นไม้การตัดสินใจ (Decision Tree) จะทำงานโดยสร้างกฎในรูปแบบโครงสร้างต้นไม้ โดยจะสร้างต้นไม้จากบนลงล่างแบบวนซ้ำ (Recursive) ซึ่งรูปแบบของต้นไม้จะเริ่มจาก Root Node ซึ่งก็คือ Decision Node แรก จาก Root Node ก็จะแตกออกเป็นโหนดลูกซึ่งโหนดลูกก็จะมีการตัดสินใจในแต่ละกิ่งหรือโหนดของตัวเอง และแต่ละโหนดจะแสดงข้อมูลที่ได้แบ่งแยกเป็นประเภทหรือข้อมูลที่ได้ตัดสินใจแล้ว โดยจะทำการแตกกิ่งย่อยออกไปตามเส้นทางของโครงสร้างของ Tree จนกระทั่งถึงโหนดในระดับสุดท้ายที่เรียกว่า โหนดปลายทาง (Leaf Node) ดังรูปที่ 2.2



รูปที่ 2.2 โครงสร้างของ Decision Tree

2.4.2 อัลกอริทึม ID3

เป็นอัลกอริทึมในการสร้างต้นไม้ตัดสินใจ ที่ใช้หลักการทฤษฎีข่าวสาร (Information Theory) คือ ค่าสารสนเทศของข้อมูลจะขึ้นอยู่กับความน่าจะเป็นของข้อมูล ค่าที่วัดได้ เรียกว่า ค่าสารสนเทศของข้อมูล (Information Measure) จะนำมาใช้ตัดสินใจว่าจะใช้ตัวแปรใดในการทำนายหรือแบ่งกลุ่มประเภทของข้อมูล โดยวิธีการกำหนดโครงสร้างต้นไม้การตัดสินใจสำหรับอัลกอริทึม ID3 นั้น จะใช้ค่า Information Gain สูงสุดในการตัดสินใจเลือก Attribute ที่จะนำมาสร้างเป็น Root Node โดยจะทำการคำนวณหาค่า Information Gain ของทุก ๆ Attributes แล้ว Attribute ที่มีค่า Gain สูงสุดจะถูกเลือกให้เป็น Decision Node แรก (Root Node) จากนั้นจะทำการแตกข้อมูลไปตามกิ่งต่างๆ (Branch) ของ Root Node จนกว่าจะจัดกลุ่มของข้อมูลได้ครบทุกกลุ่มจนถึงโหนดในระดับสุดท้ายที่เรียกว่า โหนดปลายทาง (Leaf Node) ซึ่งถ้าแต่ละกิ่งข้อมูลยังไม่เป็นกลุ่มเดียวกันหรือยังมี Attribute ที่เหลือให้เลือกอีก ก็จะทำการวนซ้ำสร้าง Node ตามหลักการเดิมเพื่อแบ่งข้อมูลไปเรื่อยๆ โดย Attribute ที่ถูกเลือกนั้นจะใช้เป็น Attribute ในการทดสอบเพื่อหา Attribute ที่มีคุณสมบัติดีที่สุดมาใช้ในการแบ่งกลุ่มข้อมูลต่อไปจนได้ต้นไม้ที่สมบูรณ์

2.4.2.1 สมการของอัลกอริทึม ID3

อัลกอริทึม ID3 จะใช้ค่า Information Gain สูงสุดในการตัดสินใจเลือก Attribute ที่จะนำมาสร้างเป็น Root Node หาได้จากสมการดังนี้

ค่าคาดคะเนของข้อมูล (Entropy) เป็นการวัดค่า Information ของข้อมูลกลุ่มตัวอย่าง (Training data set) เพื่อใช้ในการหาความเป็นไปได้ในการแยก Class ของ Target Attribute เพื่อสร้างเงื่อนไขการแบ่งแยกข้อมูลในต้นไม้ตัดสินใจ โดยข้อมูลที่ถูกเก็บจะถูกแบ่งออกเป็น c Class ดังสมการที่ 2.1

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i \qquad (2.1)$$

โดยที่

S คือ ชุดข้อมูลใดๆ ของกลุ่มตัวอย่าง (Training data set) หรือ Attributes

c คือ ข้อมูลที่ค่าแตกต่างกัน แบ่งเป็น c Class

p_i คือ สัดส่วนของจำนวนสมาชิกของกลุ่ม i กับจำนวนสมาชิกทั้งหมดใน S

ค่า Information Gain คือ เป็นค่าที่ใช้ในการเลือก Attributes มาใช้ในการแบ่งแยกข้อมูลและใช้ในการทำ Condition เพื่อสร้างต้นไม้ในขั้นตอนต่อไป ซึ่งการหาค่า Information Gain จะทำการหาค่าทุกๆ Attributes ที่ไม่ใช่ Target Attribute สามารถหาได้จากสมการที่ 2.2

$$Gain(S, A) = E(S) - \sum_{v \in value(A)} \frac{|S_v|}{|S|} E(S_v) \qquad (2.2)$$

โดยที่

A คือ Attributes

$Value(A)$ คือ กลุ่มของค่าที่เป็นไปได้ทั้งหมดของ Attribute A

$|S_v|$ คือ Subset ของ S ของ Attribute A ที่เป็นค่าของ v

$|S|$ คือ จำนวนสมาชิกของกลุ่มตัวอย่างทั้งหมด

ตารางที่ 2.1 แสดงชุดข้อมูลกลุ่มตัวอย่าง (Training data set) การตัดสินใจในการเลือกซื้อรถยนต์

Age	Income	Working	Credit	Product (Car)
<= 30	High	No	Fair	No
<= 30	High	No	Excellent	No
31-40	High	No	Fair	Yes
>40	Medium	No	Fair	Yes
>40	Low	Yes	Fair	Yes
>40	Low	Yes	Excellent	Yes
31-40	Low	Yes	Excellent	Yes
<= 30	Medium	No	Fair	No
> 40	Medium	Yes	Fair	No
<= 30	Low	Yes	Fair	Yes
<= 30	Medium	Yes	Excellent	Yes
31-40	Medium	No	Excellent	Yes
>40	Medium	No	Excellent	No
31-40	High	Yes	Fair	Yes

จากตารางที่ 2.1 เป็นชุดข้อมูลตัวอย่างการตัดสินใจในการเลือกซื้อรถยนต์ โดยมี Class labels Product (Car) เป็น Target Attribute ซึ่งมี 2 ค่า คือ Yes, No สิ่งที่นำมาพิจารณาสำหรับการตัดสินใจในการเลือกซื้อรถยนต์มี 4 Attributes คือ อายุ (Age) รายได้ (Income) สถานะการทำงาน (Working) และเครดิต (Credit) ซึ่งมีรายละเอียดดังต่อไปนี้

- อายุ (Age) แบ่งออกได้ 3 ค่า คือ น้อยกว่า 30, ระหว่าง 31-40 และมากกว่า 40 ปี
- รายได้ (Income) แบ่งออกได้ 3 ค่า คือ สูง (High), ปานกลาง (Medium), และต่ำ (Low)
- สถานะการทำงาน (Working) แบ่งออกได้ 2 ค่า คือ ทำงาน (Yes), ไม่ทำงาน (No)
- เครดิต (Credit) แบ่งออกได้ 2 ค่า คือ ดี (Excellent), ปกติ (Fair)

ดังนั้นการจำแนกประเภทข้อมูลโดยการด้วยอัลกอริทึม ID3 มีขั้นตอนดังนี้

1. คำนวณหาค่า Entropy จากตารางที่ 2.1 จำนวนข้อมูลที่อยู่ในคลาส มีค่า Yes = 9 และ No = 5 จากข้อมูลกลุ่มตัวอย่าง (Training data set) ทั้งหมด 14 ตัวอย่าง ทำการคำนวณจากสูตรในสมการที่ 2.1

จะได้

$$E(S) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$$

$$= 0.9403$$

2. คำนวณหาค่า Information Gain จากสมการที่ 2.2 โดยพิจารณาทีละAttribute ตามลำดับ คือ

- พิจารณาจาก Attribute "Age" ซึ่งมีอยู่ด้วยกัน 3 กลุ่ม ดังตารางที่ 2.2

ตารางที่ 2.2 แสดงข้อมูลสำหรับการพิจารณาที่ Attribute "Age"

Age	Sum	Yes	No
<= 30	5	2	3
31-40	4	4	0
>40	5	3	2

จากข้อมูลตารางที่ 2.2 จะได้

$$Gain(S, Age) = 0.9403 - \left(\left(\frac{5}{14}\left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right)\right) + \left(\frac{4}{14}\left(-\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4}\right)\right) + \left(\frac{5}{14}\left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right)\right)\right)$$

$$= 0.2469$$

- พิจารณาจาก Attribute "Income" ซึ่งมีอยู่ด้วยกัน 3 กลุ่ม ดังตารางที่ 2.3

ตารางที่ 2.3 แสดงข้อมูลสำหรับการพิจารณาที่ Attribute "Income"

Income	Sum	Yes	No
High	4	2	2
Medium	6	3	3
Low	4	4	0

จากข้อมูลตารางที่ 2.3 จะได้

$$Gain(S, Income) = 0.9403 - \left(\left(\frac{4}{14}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right)\right) + \left(\frac{6}{14}\left(-\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}\right)\right) + \left(\frac{4}{14}\left(-\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4}\right)\right)\right)$$

$$= 0.2260$$

- พิจารณาจาก Attribute “Working” ซึ่งมีอยู่ด้วยกัน 2 กลุ่ม ดังตารางที่ 2.4

ตารางที่ 2.4 แสดงข้อมูลสำหรับการพิจารณาที่ Attribute “Working”

Working	Sum	Yes	No
Yes	7	3	4
No	7	6	1

จากข้อมูลตารางที่ 2.4 จะได้

$$Gain(S,Working) = 0.9403 - \left(\left(\frac{7}{14}\left(-\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7}\right)\right) + \left(\frac{7}{14}\left(-\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7}\right)\right)\right)$$

$$= 0.1516$$

- พิจารณาจาก Attribute “Credit” ซึ่งมีอยู่ด้วยกัน 2 กลุ่ม ดังตารางที่ 2.5

ตารางที่ 2.5 แสดงข้อมูลสำหรับการพิจารณาที่ Attribute “Credit”

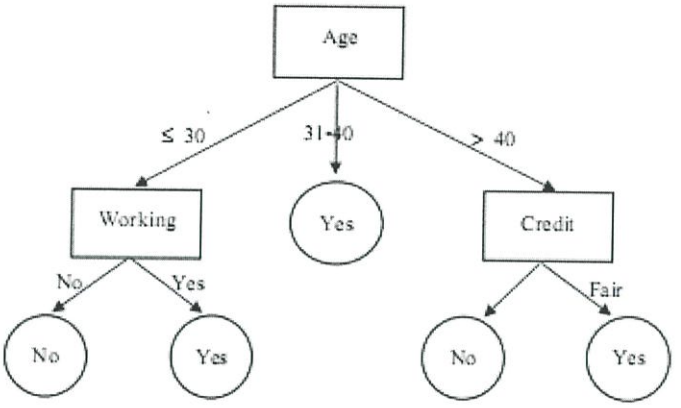
Credit	Sum	Yes	No
Excellent	8	5	3
Fair	6	4	2

จากข้อมูลตารางที่ 2.5 จะได้

$$Gain(S,Credit) = 0.9403 - \left(\left(\frac{8}{14}\left(-\frac{5}{8}\log_2\frac{5}{8} - \frac{3}{8}\log_2\frac{3}{8}\right)\right) + \left(\frac{6}{14}\left(-\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6}\right)\right)\right)$$

$$= 0.0398$$

เมื่อคำนวณหาค่า Information Gain ครบทุก Attributes จะพบว่า Gian (Age) มีค่ามากสุด ดังนั้นจึงนำมาเป็นทางเลือกตัวแรก และจะใช้เป็นเงื่อนไขในการตัดสินใจเลือก Attribute อื่นๆในการแบ่งข้อมูลลำดับถัดไป ซึ่งก็จะใช้สูตรคำนวณเช่นเดียวกัน และก็จะวนทำซ้ำไปจนกว่าข้อมูลที่มีจะหมดไป หรือแต่ละกิ่งของข้อมูลเป็นกลุ่มเดียวกัน จากการคำนวณสามารถสรุปผลได้ ดังรูปที่ 2.3



รูปที่ 2.3 แสดงต้นไม้ตัดสินใจในการซื้อรถยนต์

จากต้นไม้ตัดสินใจ สามารถเขียนในรูปแบบเงื่อนไข (Condition) ดังนี้

IF Age = <=30 AND Working = No THEN Product = "No"

IF Age = <=30 AND Working = Yes THEN Product = "Yes"

IF Age = 31-40 AND THEN Product = "Yes"

IF Age = >40 AND Credit = Excellent THEN Product = "No"

IF Age = <=30 AND Credit = Fair THEN Product = "Yes"

2.4.3 อัลกอริทึม C4.5

อัลกอริทึม C4.5 พัฒนาต่อมาจาก ID3 ซึ่งในอัลกอริทึม ID3 จะใช้ค่ามาตรฐาน Gain เป็นหลักในการเลือก Attribute ที่จะใช้เป็น (Root Node) ของต้นไม้ตัดสินใจหรือของต้นไม้ย่อย ซึ่งค่ามาตรฐาน Gain จะมีอคติ (Bias) มากกับข้อมูลที่ประกอบด้วย Attribute ที่มีค่าเป็นไปได้จำนวนมากๆ เช่น หมายเลขประจำตัวซึ่งปกติจะมีค่าไม่ซ้ำกัน ถ้าแบ่งข้อมูลตาม Attribute นี้จะทำให้เกิด Subset จำนวนมากซึ่งแต่ละ Subset จะประกอบด้วยข้อมูลเพียง 1 record ต่อกิ่งของต้นไม้ เมื่อคำนวณหาค่า Entropy จากการแบ่งข้อมูลตาม Attribute เหล่านี้ จะทำให้ค่ามาตรฐาน Gain ของ Attribute นั้นมีค่าสูงมาก ซึ่งไม่สามารถนำมาใช้เป็น Node ของต้นไม้ได้ ดังนั้นในอัลกอริทึม C4.5 จึงแก้ไขความ Bias ของค่ามาตรฐาน Gain นี้ โดยใช้ Gain Ratio สูงสุดในการตัดสินใจเลือก Attribute โดยมีการปรับค่ามาตรฐาน Gain ให้ถูกต้อง จากการใช้ค่าสารสนเทศการแบ่งแยก (Split Information) ของแต่ละ Attribute เพื่อใช้ในการคำนวณหาค่า Gain Ratio ในการสร้างแบบจำลองต้นไม้

2.4.3.1 สมการของอัลกอริทึม C4.5

ค่า Split Information เป็นค่า Information ที่มีการแบ่ง S ออกเป็น c Subset ตามค่าของ Attribute A ดังสมการที่ 2.3

$$SplitInfo(S, A) = -\sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|} \tag{2.3}$$

ค่า Gain Ratio เป็นการวัดการแบ่งข้อมูลโดยใช้ Attribute นั้นๆ

$$GainRatio(S, A) = \frac{Gain(S, A)}{SplitInfo(S, A)} \tag{2.4}$$

ดังนั้น การจำแนกประเภทข้อมูลด้วยอัลกอริทึม C4.5 มีขั้นตอนดังนี้

1. การทำงานขั้นตอนแรกคล้ายกับการทำงานด้วย ID3 แต่จะไม่ได้ใช้ค่า Information Gain แต่จะใช้ค่า Gain Ratio เป็นตัวแบ่งชุดข้อมูลกลุ่มตัวอย่าง (Training data set) ในการสร้างต้นไม้ตัดสินใจ

จากตารางที่ 2.1 แสดงชุดข้อมูลกลุ่มตัวอย่าง (Training data set) การตัดสินใจในการเลือกซื้อรถยนต์ โดยการคำนวณหา Information Gain จากชุดข้อมูลกลุ่มตัวอย่าง (Training data set) ข้างต้น ได้ค่า Information Gain ของแต่ละ Attribute ดังนี้

Gain(Age) = 0.2469

Gain(Income) = 0.2260

Gain(Working) = 0.1516

Gain(Credit) = 0.0398

จากการคำนวณ Attribute "Age" มีค่ามากที่สุด ดังนั้นจึงควรใช้ Attribute "Age" ในการแบ่งข้อมูล แต่ในกรณีที่ Attribute มีค่าหลายค่า คือมี Subset จำนวนมาก แต่ละ Subset มีจำนวน 1 record ทำให้ Information Gain มีค่าสูงมาก เมื่อถูกเลือกให้มาเป็น Root Node จะทำให้มีการแบ่งต้นไม้ตัดสินใจที่กว้างมากเกินไป ดังนั้นการแก้ปัญหานี้จึงใช้ค่า Gain Ratio เป็นตัวแบ่งชุดข้อมูล

2. คำนวณหาค่า Gain Ratio จากสูตรการคำนวณดังสมการที่ 2.4

จากตาราง Attribute "Age" มีทั้งหมด 3 Subset คือ Age = <=30, Age = 31-40, Age = > 40 ที่ประกอบด้วย 5, 4, 5 record ตามลำดับ ดังนั้น

$$SplitInfo(S, Age) = \left(-\frac{5}{14}\log_2\frac{5}{14}\right) + \left(-\frac{4}{14}\log_2\frac{4}{14}\right) + \left(-\frac{5}{14}\log_2\frac{5}{14}\right)$$

$$= 1.5774$$

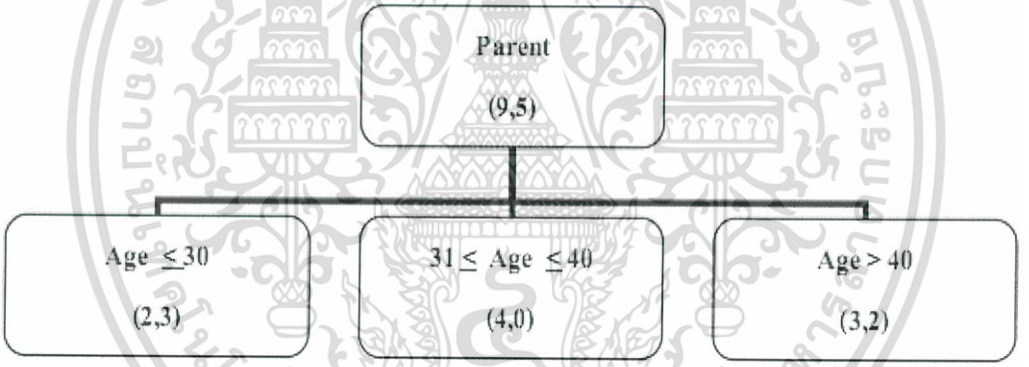
$$GainRatio(S, Age) = \frac{0.2469}{1.5774}$$

$$= 0.1565$$

ดังนั้นจะได้ค่า Gain Ratio เพื่อใช้ในการคำนวณค่า Attribute อื่นๆ ในการแบ่งชุดข้อมูลต่อไป ซึ่งวิธีนี้จะทำให้ต้นไม้ตัดสินใจมีขนาดเล็กกว่าอัลกอริทึม ID3 เพราะอัลกอริทึม ID3 มีค่า Gain ที่มีความโน้มเอียงมาก เช่น การใช้ค่า ID ซึ่งแต่ละ Record ไม่มีค่าซ้ำกันและมีจำนวนมาก โดยถ้ามีจำนวน 14 record จะต้องทำการสร้าง 14 กิ่ง โดยแต่ละกิ่งจะมีค่าเพียงอย่างเดียวระหว่าง Yes กับ No โดยแต่ละกิ่งจะมีค่า Info ([1,0]) หรือ Info ([0,1])

$$\text{Info}([1,0],[0,1]),\ldots,[1,0],[0,1]) = 0$$

โดยจะได้ค่า Information Gain = 0.94-0 = 0.94 ซึ่งมีค่าเท่าเดิม ทำให้ไม่เกิดประโยชน์ต่อการทำงาน แต่อัลกอริทึม C4.5 จะมีการทำ SplitInfo เพิ่มเข้ามาเพื่อที่จะได้ลดตัวแปรที่มีความเป็นไปได้มากเกินความจำเป็นออก เพราะจะทำให้การแบ่งกลุ่มข้อมูลหรือพยากรณ์ได้ยาก



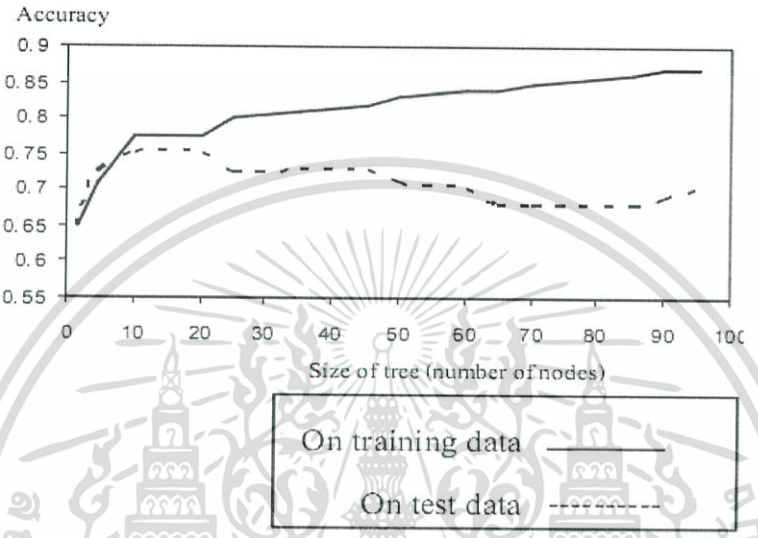
รูปที่ 2.4 แสดงการแบ่งกลุ่มข้อมูลเริ่มต้นที่ Attribute "Age"

2.4.3.2 Over-Fitting

ปัญหาในการสร้างต้นไม้ตัดสินใจ คือ Over-Fitting เป็นปัญหาที่ต้นไม้สร้าง Node และ Branch ที่มีความลึกและซับซ้อนมากเกินความจำเป็น ทำให้เกิดแนวโน้มความถูกต้องที่ต่ำลงเกิดการวิเคราะห์ข้อมูลที่ผิดพลาดได้ โดยสามารถแก้ปัญหาโดยวิธีการทำ Tree Pruning โดยมีเทคนิคอยู่ 2 วิธี คือ

1. Pre-Pruning คือ การหยุดสร้างต้นไม้หรือแตกกิ่งไม่ให้มีขนาดใหญ่เกินไป จนกว่าจะจัดกลุ่มข้อมูลได้ เช่น เมื่อตัววัดในการแตกกิ่งหรือข้อมูลที่จะแตกต่อมีจำนวนน้อยเกินไปหรือจำนวนข้อมูลใน Leaf Node มีน้อยเกินไป

2. Post-Pruning คือ เป็นการสร้าง Tree ให้แตกกิ่งไปเรื่อยๆ จนโตเต็มที่แล้วค่อยตัดกิ่งที่ไม่ดี (Unreliable) ออก



รูปที่ 2.5 ความถูกต้องจากการแยกแยะข้อมูลของต้นไม้ตัดสินใจเทียบกับขนาดของต้นไม้ตัดสินใจ

จากรูปที่ 2.5 แสดงให้เห็นว่าต้นไม้ตัดสินใจมีขนาดใหญ่ขึ้น ซึ่งจะให้ความถูกต้อง (Accurancy) กับ Trainning Data Set มากขึ้น แต่เมื่อนำไปใช้งานจริงความถูกต้องจะลดลงตามลำดับของขนาดต้นไม้ซึ่งได้จากการวัดความถูกต้องกับ Test Data

2.4.3.3 กรณีที่ข้อมูลเป็นตัวเลขต่อเนื่อง (Continuous-Value Attributes)

ในกรณีที่ Attribute มีค่าเป็นตัวเลข การทดสอบค่าของAttribute จะใช้ค่า Threshold ที่เหมาะสม โดยถ้า Z เป็นค่า Threshold ที่เหมาะสม การทดสอบค่าที่ Attribute นี้จะแบ่งเป็น $A \leq Z$ และ $A > Z$ โดยทำการเปรียบเทียบค่าของ A กับค่า Threshold (Z) โดยการหาค่า Threshold ที่เหมาะสมมีขั้นตอน คือ เรียงลำดับค่าใน Attribute A นำค่าที่อยู่ติดกันแต่ให้ผลที่มีการเปลี่ยนแปลงไปมาทำการคำนวณหากึ่งกลาง (Midpoint) ที่อยู่ในแต่ละช่วง โดยเลือกค่าที่มากที่สุด แต่ต้องไม่เกินค่ากึ่งกลาง (Midpoint) มาเป็นค่า Threshold ในแต่ละช่วง และค่า Threshold จะพิจารณาจากค่า Information Gain ที่สูงที่สุด

2.4.3.4 กรณีที่ข้อมูลขาดหายไป (Unknown Attribute Values)

เมื่อพิจารณาจาก $Gain(A)$ ที่โหนด n ใน Decision Tree เพื่อหาว่า Attribute A เป็น Attribute ที่ดีที่สุดหรือไม่ ในกรณีที่ไม่ทราบค่าของ Attribute ทำให้ไม่สามารถแบ่งกลุ่มข้อมูลได้ ซึ่งวิธีที่จะจัดการกับข้อมูลดังกล่าว มีดังนี้

1. หาค่า $Info(T)$ และ $Info_x(T)$ โดยพิจารณาจากข้อมูลที่รู้ค่าของ Attribute A
2. กำหนดค่าความน่าจะเป็นของความถี่ตามค่าใน Attribute A โดยการหาค่าจะนำมาคำนวณ Information Gain กับค่าความน่าจะเป็นของค่าที่รู้ของ Attribute A เช่น ใน Attribute A ที่ทราบค่าของ A มีจำนวน 9 record และที่ไม่ทราบมีจำนวน 1 record ค่าความน่าจะเป็นของ Attribute ที่ทราบค่ามีค่าเป็น 0.9 และนำมาคูณกับค่า Information Gain จะได้ค่าของ Information Gain ของ Attribute A สำหรับการหาค่า Split Information จะมีแบ่งเป็น subset ใหม่ สำหรับข้อมูลที่ไม่ทราบค่าใน Attribute A
3. การหาค่า Split Information จะพิจารณากลุ่มของข้อมูลที่ไม่รู้ค่าของ A เป็นอีก 1 subset เช่น Attribute ที่จะนำมาทดสอบมีค่าที่เป็นไปได้ n ค่า Split Information จะถูกคำนวณโดยแบ่งข้อมูลออกเป็น n+1 subset

การแบ่ง Training Data Set สมมติ Attribute ที่เลือกจากขั้นตอนแรกมีค่าเป็นไปได้คือ $O_1, O_2, ..., O_n$ เมื่อข้อมูล 1 record ใน T ซึ่งมี O_i ถูกกำหนดให้ subset T_i ค่าความน่าจะเป็นที่ข้อมูลนี้อยู่ใน subset T_i เท่ากับ 1 และความน่าจะเป็นที่ข้อมูลนี้อยู่ใน subset อื่นๆ เท่ากับ 0 แต่ถ้าค่าใน Attribute ที่ไม่ทราบค่า ความน่าจะเป็นจะมีค่าน้อยลง สำหรับข้อมูลแต่ละ record ในแต่ละ subset T_i weight จะเท่ากับความน่าจะเป็นของ O_i ที่จุดนั้นๆ ทำให้ T_i เป็นผลรวมค่า weight w ซึ่งค่าใน Attribute ไม่ทราบค่าจะถูกกำหนดให้แต่ละ subset T_i ด้วย weight ดังสมการ

$$\text{W x Probabillity of outcome } O_i \tag{2.5}$$

โดยความน่าจะเป็นคือ ผลรวมของ weight ของข้อมูลทั้งหมดใน T ซึ่งมีค่า O_i หารด้วยผลรวมของ weight ของข้อมูลทั้งหมดใน T ซึ่งมีค่าใน Attribute เป็นค่าที่ไม่ทราบค่า

การใช้ผลลัพธ์ต้นไม้การตัดสินใจ (Decision Tree) ที่ได้มาทำนายกลุ่มข้อมูล ในกรณีที่ค่าใน Attribute ที่จะทดสอบที่ Decision node เป็นค่าที่ไม่ทราบค่า ทำให้ไม่สามารถแบ่งข้อมูลได้ และรวมผลที่ได้จากการจำแนกประเภทด้วยวิธีการทางคณิตศาสตร์ โดยผลที่ได้จะเกิดได้หลายเส้นทางจาก Root ของ Tree หรือ ไปยัง Leaf Node และ Class ที่ได้จากการทำนายจะเป็น Class ที่มีความน่าจะเป็นสูงสุด

2.4.3.5 Tree-Pruning

อัลกอริทึม C4.5 นี้ จะทำการ Pruning โดยการตัดกิ่งที่ทำให้เกิดความผิดพลาดในการทำนายออกไป แล้วแทนที่กิ่งนั้นด้วย Leaf Node โดยเทคนิคนี้จะใช้ Tranning Data Set สำหรับสร้าง Tree และตัดกิ่งต้นไม้ตัดสินใจเท่านั้น โดยไม่ต้องใช้ชุดข้อมูลที่แยกออกต่างหากสำหรับการตัดกิ่งโดยเฉพาะ และการคำนวณความผิดพลาดหรือการประมาณค่าความผิดพลาดที่เกิดจากทำนายของแต่ละ Leaf Node และแต่ละกิ่งจะทำได้ โดยจะทำการแบ่งกลุ่ม Set สำหรับข้อมูลที่ไม่เคยพบมาก่อนที่มีขนาดเท่ากับ Tranning Data Set โดยการคำนวณจะใช้ฟังก์ชั่นทางสถิติ ซึ่งอยู่บนพื้นฐานการกระจายแบบ binomial distribution ที่ระดับความเป็นอิสระเท่ากับ CF (confidence level) คือ ถ้ามีข้อมูล N ตัวที่ Node และมีข้อมูล E ตัวเป็นข้อมูลที่มีกลุ่มไม่ถูกต้องหรือไม่ตรงกับกลุ่มส่วนใหญ่ ค่าความผิดพลาดที่ Node นี้จะสามารถเขียนได้ในรูป Ucf(E,N)

ถ้า Leaf Node ประกอบด้วยข้อมูลจำนวน N ตัว ค่าความผิดพลาดที่คาดไว้ของข้อมูลแต่ละตัวเท่ากับ N x Ucf(E,N) ตัว ดังสมการ

$$N = N \times Ucf(E,N) \tag{2.6}$$

กำหนดให้

N = ขนาดของข้อมูลที่ Leaf Node ใดๆ

E = จำนวนของ error ที่เกิดขึ้นในเซตของข้อมูลที่ Leaf Node ใดๆ

Ucf(E,N) = ความน่าจะเป็นสูงสุดที่เกิด error

ดังนั้น เมื่อทดสอบกับข้อมูลที่ไม่เคยพบมาก่อน ซึ่งถ้าค่าที่ได้จากการคำนวณจำนวนข้อมูลที่ error ของแต่ละกิ่งรวมกันแล้วมากกว่าจำนวนข้อมูลที่ error ของแต่ละกิ่งรวมกันแล้วมากกว่าจำนวนข้อมูลที่ error ของ Node ที่แตกกิ่งนั้น ก็จะตัด Node ที่เป็นลูกในทุกกิ่งของ Node ที่แตกกิ่งนั้นออกมาให้หมดจนเหลือเฉพาะ Node ที่แตกกิ่งนั้นไว้ Node เดียว

บทที่ 3
การวิเคราะห์และออกแบบ

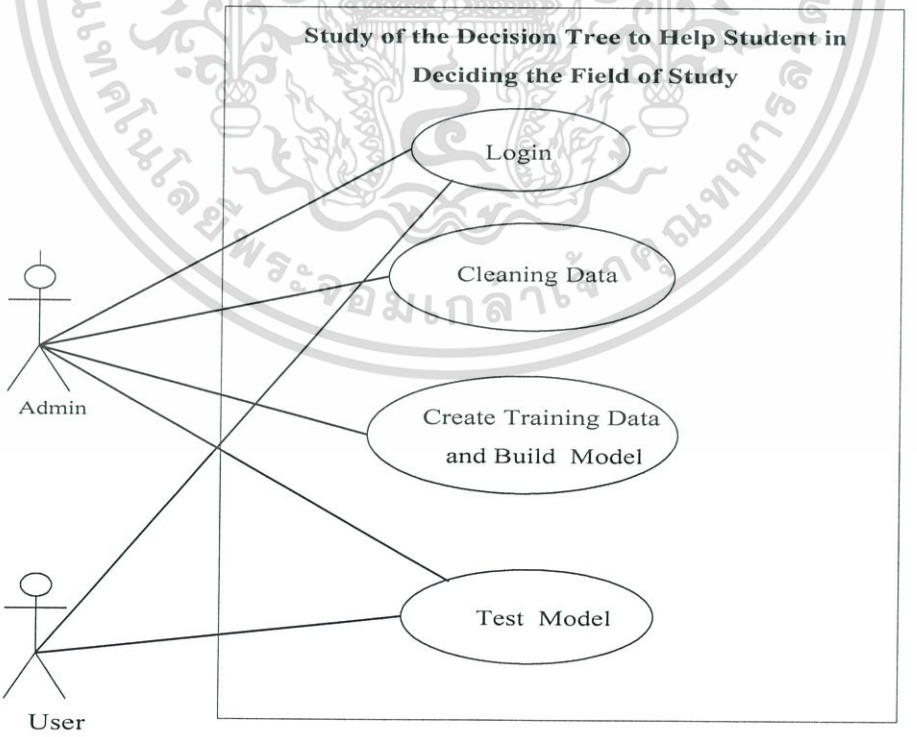
ในบทนี้จะกล่าวถึงการวิเคราะห์และออกแบบระบบ ซึ่งเป็นการเก็บรวบรวมข้อมูลเพื่อให้ทราบถึงขั้นตอนการทำงาน ปัญหาที่เกิดขึ้นในการทำงานในปัจจุบัน เพื่อเป็นข้อมูลที่จะนำมาออกแบบระบบงานให้สามารถทำงานได้ตามที่ต้องการ โดยโครงงานนี้จะเป็นการออกแบบเพื่อสร้างโมเดลต้นไม้ตัดสินใจในการแบ่งประเภทข้อมูลออกเป็นกลุ่ม ซึ่งแสดงออกมาในรูปแบบโครงสร้างต้นไม้ เพื่อช่วยในการตัดสินใจในการเลือกสาขาของนักเรียนต่อไป

3.1 การวิเคราะห์และออกแบบระบบ

การวิเคราะห์ และออกแบบแบบจำลองของระบบนั้น ทำโดยใช้ภาษา UML ซึ่งจะแสดงด้วย Diagaram แบบต่างๆ โดยการวิเคราะห์ และออกแบบจำลองของระบบงานนี้จะแสดงด้วย Use-Case Diagram, Activity Diagram, Class Diagram, Sequence Diagram

3.1.1 Use-Case Diagram

เป็นแบบจำลองการทำงานของระบบ โดยอธิบายการทำงานของระบบได้ดังนี้



รูปที่ 3.1 แสดง Use Case ของระบบ

จากรูป 3.1 เป็นการอธิบายภาพรวมของระบบว่าระบบมีการทำงานอะไรบ้าง โดยมีองค์ประกอบ 2 ส่วน คือ Use Case และ Actor โดยที่ Use Case จะแสดงถึงขอบเขตของระบบงาน ส่วน Actor คือ สิ่งที่อยู่นอกระบบซึ่งจะเป็นผู้กระตุ้นให้ระบบเกิดการทำงานหรือรับผลลัพธ์จากการกระทำของระบบด้วย โดยแผนภาพ Use Case ของการศึกษาเทคนิคของต้นไม้ตัดสินใจเพื่อช่วยตัดสินใจในการเลือกสาขาของนักเรียน ประกอบด้วย Use Case และ Actor มีรายละเอียดดังนี้

Actor ที่เกี่ยวข้องกับระบบ มีดังนี้

1. Admin คือ ผู้ที่ทำหน้าที่วิเคราะห์ระบบ

2. User คือ ผู้ใช้งานระบบ

Use Case ที่เกี่ยวข้องกับระบบ มีดังนี้

1. Use Case : Login เป็นการทำงานในส่วนของการตรวจสอบความการเข้าใช้งานระบบ

2. Use Case : Cleaning Data เป็นการทำงานในส่วนของการตรวจสอบความถูกต้องของข้อมูลเพื่อใช้ในการทำ Classification

3. Use Case : Create Training Data and Build Model เป็นการทำงานในส่วนการสร้างชุดข้อมูลตัวอย่าง (Training Data Set) ที่ได้ทำการจัดรูปแบบข้อมูลที่มีความถูกต้องเหมาะสม และเป็นการทำงานในส่วนของการสร้างโมเดลต้นไม้

4. Use Case :Test Model เป็นการทำงานในส่วนของการทดสอบโมเดลหรือจัดแบ่งประเภทข้อมูลที่ได้จากการสร้างโมเดลต้นไม้

3.1.2 Use-Case Description

เป็นคำอธิบายรายละเอียดการทำงานของแต่ละ Use-Case ว่ามีการทำงานอย่างไรหรือการทำงานในส่วนใดที่มีความสัมพันธ์กันบ้าง แสดงได้ดังนี้

ตารางที่ 3.1 แสดงรายละเอียดการทำงานของ Use-Case Login

<table>
<tr><td>Use-Case Name</td><td colspan="2">Login</td></tr>
<tr><td>Scenario</td><td colspan="2">ตรวจสอบผู้ใช้งานระบบ</td></tr>
<tr><td>Triggering Event</td><td colspan="2">ต้องการตรวจสอบผู้ใช้งานระบบ</td></tr>
<tr><td>Brief Description</td><td colspan="2">การเข้าใช้งานระบบจะต้องมีการ Login โดยมีการป้อน Username, Password</td></tr>
<tr><td>Actor</td><td colspan="2">Admin, User</td></tr>
<tr><td>Stakeholders</td><td colspan="2">-</td></tr>
<tr><td>Preconditions</td><td colspan="2">-</td></tr>
<tr><td>Post conditions</td><td colspan="2">สามารถเข้าใช้งานระบบได้</td></tr>
<tr><td rowspan="2">Flow of Activity</td><td>Actor</td><td>System</td></tr>
<tr><td>1. เข้าสู่หน้าการตรวจสอบผู้ใช้งานระบบ
2. ผู้ใช้กรอก Username, Password
3. กดปุ่มยืนยันการเข้าใช้งานระบบ</td><td>1. แสดงหน้าการตรวจสอบผู้ใช้งานระบบ
3. ระบบจะทำการตรวจสอบสิทธิ์ของผู้ใช้งาน
3.1 ถ้าผู้เข้าใช้งานมีสิทธิ์ ระบบจะแสดงหน้าหลักในการใช้งาน
3.2 ถ้าผู้เข้าใช้งานไม่มีสิทธิ์ ระบบจะทำการให้กรอกข้อมูลของผู้ใช้งานอีกครั้ง</td></tr>
<tr><td>Exception Conditions</td><td colspan="2">ถ้า Username, Password ไม่ถูกต้องระบบจะทำการให้กรอกข้อมูลใหม่</td></tr>
</table>

ตารางที่ 3.2 แสดงรายละเอียดการทำงานของ Use-Case Cleaning Data

<table>
<tr><td>Use-Case Name</td><td colspan="2">Cleaning Data</td></tr>
<tr><td>Scenario</td><td colspan="2">ตรวจสอบความถูกต้องของข้อมูลเพื่อใช้ในการทำ Classification</td></tr>
<tr><td>Triggering Event</td><td colspan="2">ต้องการตรวจสอบความถูกต้องของข้อมูล</td></tr>
<tr><td>Brief Description</td><td colspan="2">ตรวจสอบความถูกต้องของข้อมูล และทำการแก้ไขข้อมูลที่ผิดปกติให้อยู่ในรูปแบบที่เหมาะสม ก่อนที่จะนำข้อมูลนั้นไปใช้ในการทำ Classification ต่อไป</td></tr>
<tr><td>Actor</td><td colspan="2">Admin</td></tr>
<tr><td>Stakeholders</td><td colspan="2">-</td></tr>
<tr><td>Preconditions</td><td colspan="2">-</td></tr>
<tr><td>Post conditions</td><td colspan="2">ข้อมูลมีความถูกต้อง และอยู่ในรูปแบบที่เหมาะสม</td></tr>
<tr><td>Flow of Activity</td><td>Actor</td><td>System</td></tr>
<tr><td></td><td>1. เข้าสู่หน้าการเตรียมข้อมูล (Models)
2. เลือกปุ่ม “Create new models”
3. เลือกปุ่ม “Check Data”
4. เลือกวิธีการในการแก้ไขข้อมูล</td><td>1. แสดงหน้าการเตรียมข้อมูล (Models)
2. แสดงหน้า Create new models
3. ระบบจะทำการตรวจสอบข้อมูล
3.1 ข้อมูลถูกต้อง ระบบจะแสดง “ Data Complete”
3.2 ข้อมูลผิดปกติ ระบบจะแสดงแบบฟอร์มในการแก้ไขข้อมูล
4. ระบบจะทำการแก้ไขข้อมูลให้ถูกต้องตามวิธีที่เลือก และทำการบันทึกข้อมูลเข้าในระบบ และแสดง “ Data Complete”</td></tr>
<tr><td>Exception Conditions</td><td colspan="2">-</td></tr>
</table>

ตารางที่ 3.3 แสดงรายละเอียดการทำงานของ Use-Case Create Training Data and Build Model

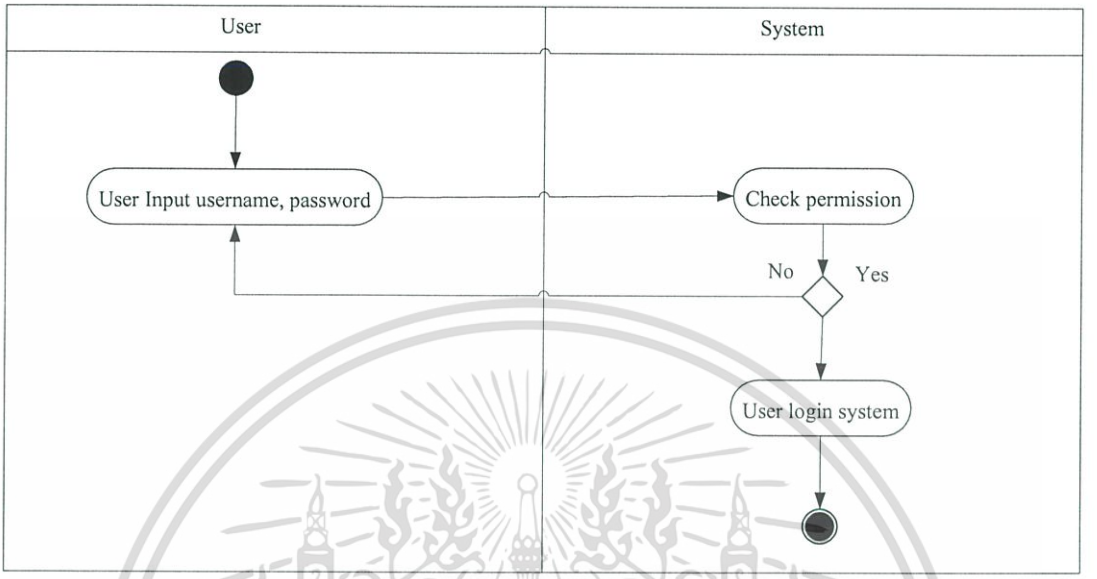
<table>
<tr><td>Use-Case Name</td><td colspan="2">Create Training Data and Build Model</td></tr>
<tr><td>Scenario</td><td colspan="2">สร้างชุดข้อมูลตัวอย่าง (Training Data Set) และสร้างโมเดลต้นไม้ตัดสินใจ</td></tr>
<tr><td>Triggering Event</td><td colspan="2">ทำการสร้างชุดข้อมูลตัวอย่าง (Training Data Set) และสร้างโมเดลต้นไม้ตัดสินใจ</td></tr>
<tr><td>Brief Description</td><td colspan="2">สร้างชุดข้อมูลตัวอย่าง (Training Data Set) ที่ได้ทำการจัดรูปแบบของข้อมูลให้อยู่ในรูปแบบที่ถูกต้องเหมาะสม และสร้างโมเดลต้นไม้ตัดสินใจ เพื่อนำไปใช้ในการวิเคราะห์ข้อมูลใหม่ๆ ที่เข้ามา</td></tr>
<tr><td>Actor</td><td colspan="2">Admin</td></tr>
<tr><td>Stakeholders</td><td colspan="2">-</td></tr>
<tr><td>Preconditions</td><td colspan="2">ในกรณีสร้างโมเดลต้นไม้ตัดสินใจ ต้องสร้างชุดข้อมูลตัวอย่าง (Training Data Set) และ Target Attribute ที่ต้องการแบ่งกลุ่มประเภทข้อมูล</td></tr>
<tr><td>Post conditions</td><td colspan="2">ได้โมเดลต้นไม้ตัดสินใจ</td></tr>
<tr><td rowspan="5">Flow of Activity</td><td>Actor</td><td>System</td></tr>
<tr><td>1. เข้าสู่หน้า Models</td><td>1. แสดงหน้า Models</td></tr>
<tr><td>2. เลือกปุ่ม “Create new models”</td><td>2. แสดงหน้า Create new models</td></tr>
<tr><td>3. เลือกอัลกอริทึม Classifier Tree (ID3,C4.5)</td><td></td></tr>
<tr><td>4. กดปุ่ม “Training” ยืนยันการสร้างโมเดล</td><td>4. ทำการสร้างโมเดล และแสดงผลในรูปแบบของต้นไม้ตัดสินใจ</td></tr>
<tr><td>Exception Conditions</td><td colspan="2">ทุกค่าของข้อมูลอยู่ในคลาสเดียวกัน
กำหนด Node ของต้นไม้ด้วย Target Attribute</td></tr>
</table>

ตารางที่ 3.4 แสดงรายละเอียดการทำงานของ Use-Case Test Model

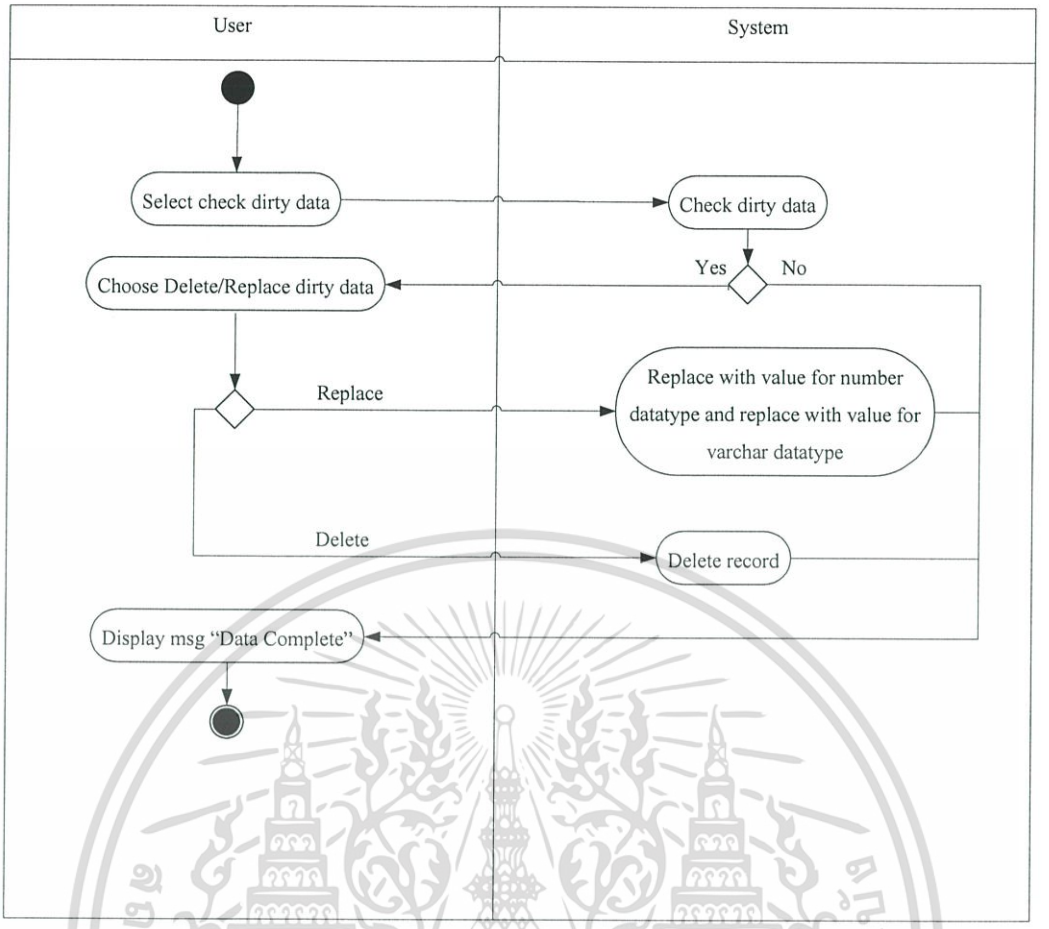
Use-Case Name	Test Model	
Scenario	ทดสอบโมเดลหรือจัดแบ่งประเภทข้อมูลที่ได้จากการสร้างโมเดลต้นไม้	
Triggering Event	ทำการทดสอบโมเดล	
Brief Description	ทำการทดสอบความถูกต้องของโมเดล โดยการใช้ข้อมูลทดสอบ (Test Data) ซึ่งเป็นคนละชุดกับที่ใช้ในการสร้างชุดข้อมูลตัวอย่าง (Training Data Set) และสามารถทำนายข้อมูลที่ผู้ใช้ทำการกรอกข้อมูลใหม่เข้าไปโดยใช้โมเดลที่เลือก	
Actor	Admin, User	
Stakeholders	-	
Preconditions	ต้องผ่านขั้นตอนการสร้างโมเดลก่อน	
Post conditions	ได้โมเดลต้นไม้ตัดสินใจ	
Flow of Activity	**Actor**	**System**
	1. เข้าสู่หน้าการ Test Data 2. เลือกอัลกอริทึม (ID3,C4.5) ที่ใช้ในการทดสอบ 3. เลือกข้อมูลที่ต้องการนำมาทดสอบโมเดล 3.1 เลือกข้อมูลการทดสอบโดยการนำเข้าข้อมูลจากไฟล์ข้อมูล 3.2 เลือกข้อมูลการทดสอบโดยการกรอกผลการเรียนเฉลี่ย, ผลการเรียนเฉลี่ยแต่ละวิชา, เพศ, สาขาวิชาที่ต้องการทำนาย, 4. กดปุ่ม “Test” ยืนยันการทดสอบโมเดล	1. แสดงหน้า Test Data 4. ทำการทดสอบโมเดลและแสดงผลการทดสอบโมเดลที่ได้จากต้นไม้ตัดสินใจ
Exception Conditions	แสดงข้อความเตือน “Can’t test data” หากข้อมูลไม่ Map กับข้อมูลของโมเดล	

3.1.3 Activity Diagram

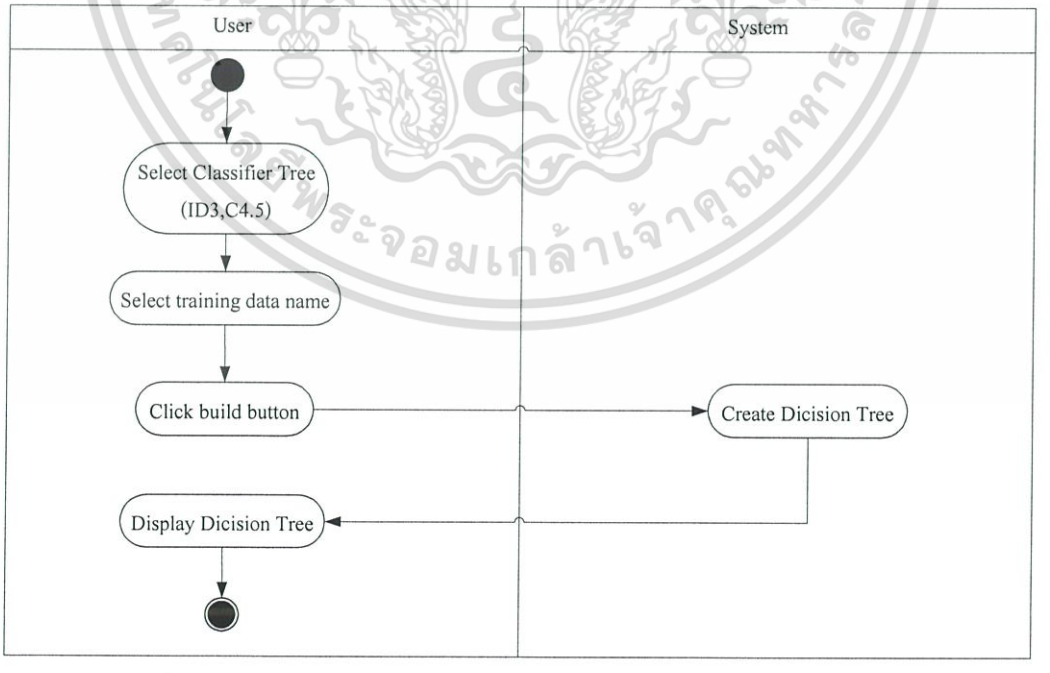
ใช้อธิบายรายละเอียดขั้นตอนการทำงานของระบบในแต่ละ Use-Case โดยที่ขั้นตอนการทำงานในแต่ละขั้นตอน เรียกว่า Activity โดยรายละเอียดสามารถแสดงได้ ดังนี้



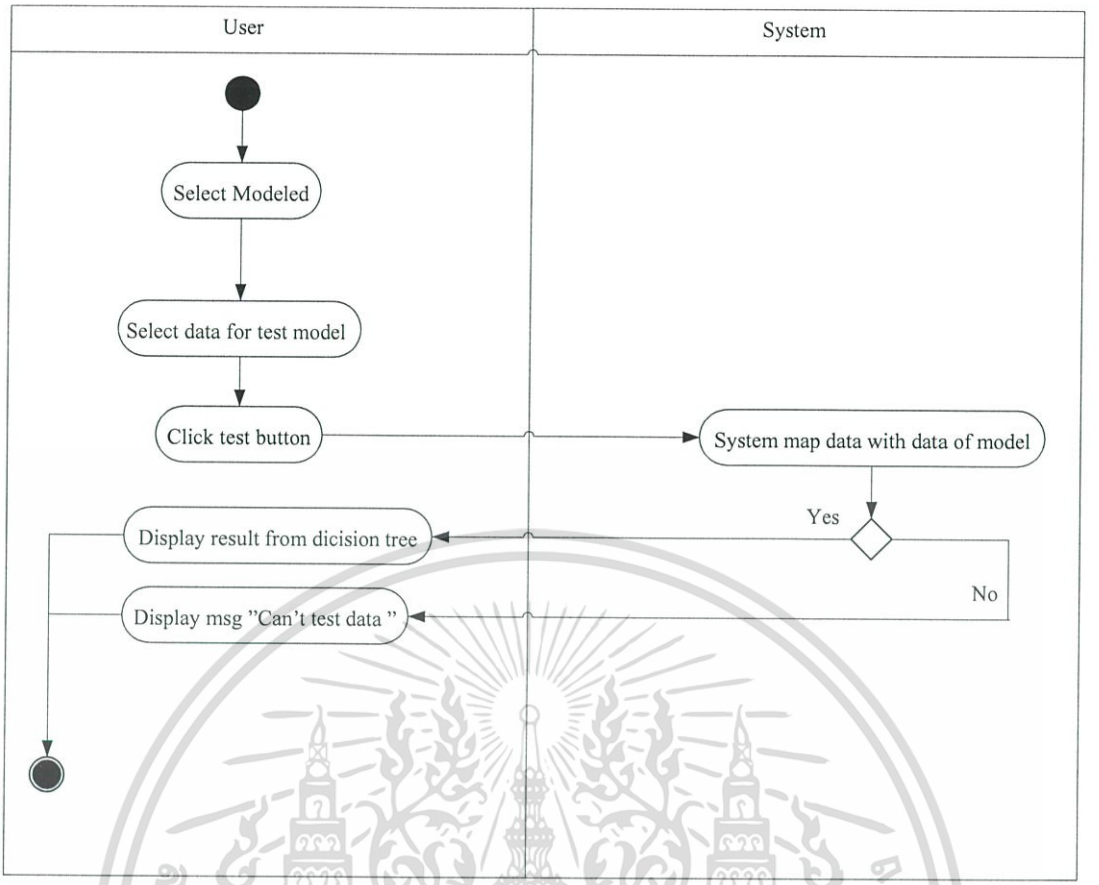
รูปที่ 3.2 แสดง Activity Diagram ของ Login



รูปที่ 3.3 แสดง Activity Diagram ของ Cleaning Data

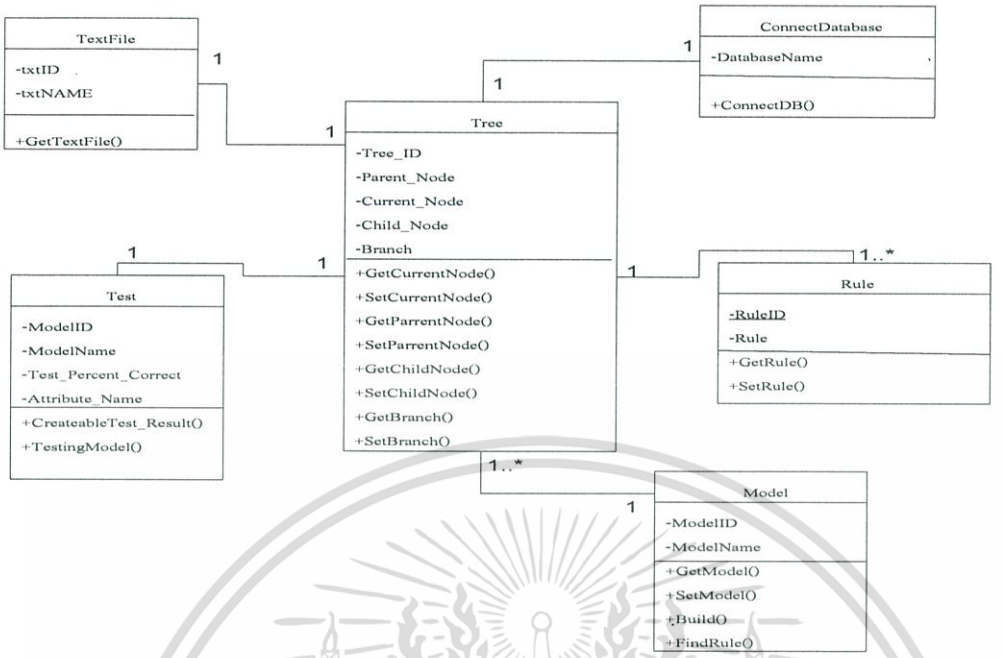


รูปที่ 3.4 แสดง Activity Diagram ของ Create Training Data and Build Model



รูปที่ 3.5 แสดง Activity Diagram ของ Test Model

3.1.4 Class Diagram

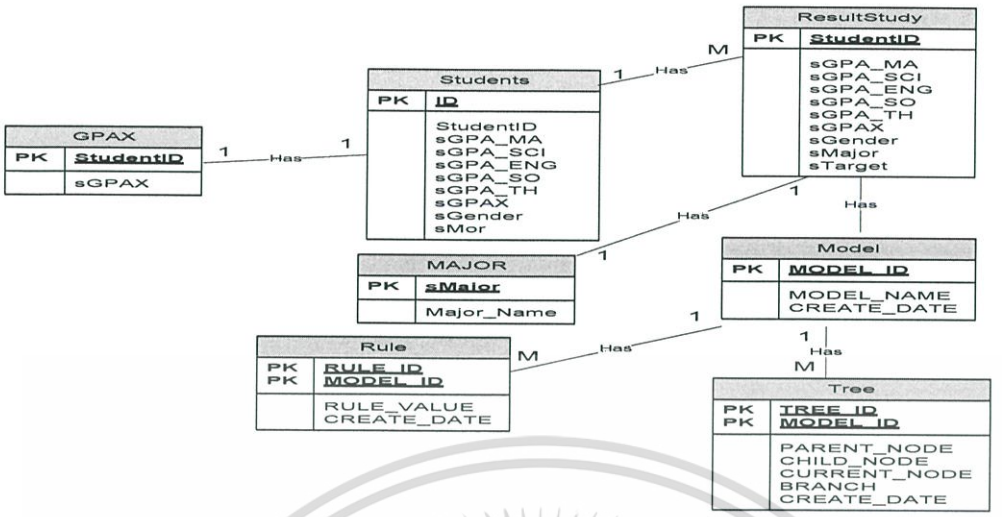


รูปที่ 3.6 แสดง Class Diagram ของระบบการศึกษาเทคนิคของต้นไม้ตัดสินใจเพื่อช่วยตัดสินใจในการเลือกสาขาของนักเรียน

จากรูปที่ 3.6 Class Diagram ของระบบการศึกษาเทคนิคของต้นไม้ตัดสินใจเพื่อช่วยตัดสินใจในการเลือกสาขาของนักเรียน มีดังนี้

- คลาส Tree เป็นคลาสที่เก็บข้อมูลของ Tree ที่ได้จากการสร้างโมเดลต้นไม้ตัดสินใจ
- คลาส TextFile เป็นคลาสที่เก็บข้อมูลของ Tree จากแต่ละอัลกอริทึมที่เลือกใช้
- คลาส Test เป็นคลาสที่ใช้ในการทดสอบโมเดลที่ได้จากการสร้างโมเดล Tree
- คลาส Rule เป็นคลาสที่เก็บข้อมูลของกฎที่ได้จากการสร้างโมเดล Tree
- คลาส Model เป็นคลาสที่เก็บข้อมูลเกี่ยวกับชื่อโมเดลที่สร้าง

3.1.5 ER Diagram



รูปที่ 3.7 แสดง ER Diagram ของระบบการศึกษาเทคนิคของต้นไม้ตัดสินใจเพื่อช่วยตัดสินใจในการเลือกสาขาของนักเรียน

จากรูปที่ 3.7 ER Diagram ของระบบการศึกษาเทคนิคของต้นไม้ตัดสินใจเพื่อช่วยตัดสินใจในการเลือกสาขาของนักเรียน มีดังนี้

- เอนทิตี Students คือ เอนทิตีสำหรับจัดเก็บข้อมูลนักเรียน ประกอบด้วยแอตทริบิวต์ต่างๆ เช่น รหัสนักเรียน, เพศ, สาขาการเรียน, คะแนนกลุ่มวิชาต่างๆ เป็นต้น
- เอนทิตี Major คือ เอนทิตีสำหรับจัดเก็บข้อมูลสาขาการเรียน ประกอบด้วยแอตทริบิวต์ต่างๆ เช่น รหัสสาขาการเรียน, ชื่อสาขาการเรียน เป็นต้น
- เอนทิตี ResultStudy คือ เอนทิตีสำหรับจัดเก็บข้อมูลผลการเรียน ประกอบด้วยแอตทริบิวต์ต่างๆ เช่น รหัสนักเรียน, คะแนนเฉลี่ยกลุ่มวิชาต่างๆ, เพศ, สาขาการเรียน เป็นต้น
- เอนทิตี GPAX คือ เอนทิตีสำหรับจัดเก็บข้อมูลผลการเรียนเฉลี่ยตลอดหลักสูตร ประกอบด้วยแอตทริบิวต์ต่างๆ เช่น รหัสนักเรียน, คะแนนเฉลี่ยตลอดหลักสูตร เป็นต้น
- เอนทิตี Rule คือ เอนทิตีสำหรับจัดเก็บข้อมูลของกฎที่ได้จากการสร้างโมเดล Tree ประกอบด้วยแอตทริบิวต์ต่างๆ เช่น รหัสกฎ, รหัสโมเดล เป็นต้น
- เอนทิตี Model คือ เอนทิตีสำหรับจัดเก็บรายละเอียดของโมเดลที่ได้ทำการสร้าง ประกอบด้วยแอตทริบิวต์ต่างๆ เช่น รหัสโมเดล, ชื่อโมเดล เป็นต้น
- เอนทิตี Tree คือ เอนทิตีสำหรับจัดเก็บข้อมูลข้อมูลของ Tree ที่ได้จากการสร้างโมเดลต้นไม้ตัดสินใจ ประกอบด้วยแอตทริบิวต์ต่างๆ เช่น รหัสทรี, ชื่อโมเดล, โหนดต่างๆ เป็นต้น

3.1.6 พจนานุกรมข้อมูล

จาก ER Diagram ของระบบการศึกษาเทคนิคของต้นไม้ตัดสินใจเพื่อช่วยตัดสินใจในการเลือกสาขาของนักเรียน สามารถแสดงรายละเอียดในพจนานุกรมได้ ดังนี้

ตารางที่ 3.5 แสดงข้อมูล Entity ของนักเรียน (Students)

Attribute	Type	Detail	Key	Reference
ID	int	รหัส	PK	
StudentID	varchar (50)	รหัสนักเรียน		
sGPA_MA	float	คะแนนกลุ่มวิชาคณิตศาสตร์		
sGPA_SCI	float	คะแนนกลุ่มวิชาวิทยาศาสตร์		
sGPA_ENG	float	คะแนนกลุ่มวิชาอังกฤษ		
sGPA_SO	float	คะแนนกลุ่มวิชาสังคม		
sGPA_TH	float	คะแนนกลุ่มวิชาภาษาไทย		
sGPAX	float	คะแนนตลอดหลักสูตร		
sGender	float	เพศ		
sMor	float	ระดับชั้นมัธยม1-3		

ตารางที่ 3.6 แสดงข้อมูล Entity ของสาขาการเรียน (Major)

Attribute	Type	Detail	Key	Reference
sMajor	varchar(3)	รหัสสาขาการเรียน	PK	
Major_Name	varchar (50)	ชื่อสาขาการเรียน		

ตารางที่ 3.7 แสดงข้อมูล Entity ของคะแนนเฉลี่ยตลอดหลักสูตร (GPAX)

Attribute	Type	Detail	Key	Reference
StudentID	varchar (50)	รหัสนักเรียน	PK	
sGPAX	float	คะแนนเฉลี่ยตลอดหลักสูตร		

ตารางที่ 3.8 แสดงข้อมูล Entity ของผลการเรียน ResultStudy

Attribute	Type	Detail	Key	Reference
StudentID	varchar (50)	รหัสนักเรียน	PK	
sGPA_MA	float	คะแนนเฉลี่ยกลุ่มวิชาคณิตศาสตร์		
sGPA_SCI	float	คะแนนเฉลี่ยกลุ่มวิชาวิทยาศาสตร์		
sGPA_ENG	float	คะแนนเฉลี่ยกลุ่มวิชาอังกฤษ		
sGPA_SO	float	คะแนนเฉลี่ยกลุ่มวิชาสังคม		
sGPA_TH	float	คะแนนเฉลี่ยกลุ่มวิชาภาษาไทย		
sGPAX	float	คะแนนเฉลี่ยตลอดหลักสูตร		
sGender	float	เพศ		
sMajor	float	สาขาการเรียน		
sTarget	float	ผลลัพธ์เป้าหมาย		

ตารางที่ 3.9 แสดงข้อมูล Entity ของสิทธิผู้ใช้งาน (Accounts)

Attribute	Type	Detail	Key	Reference
ID	int	รหัสลำดับผู้ใช้งาน	PK	
sLogin	varchar (50)	รหัสผู้ใช้งาน		
sPasswd	varchar (50)	รหัสผ่าน		
sLevel	int	สถานะผู้ใช้งาน		

ตารางที่ 3.10 แสดงข้อมูล Entity ของโมเดล (Model)

Attribute	Type	Detail	Key	Reference
MODEL_ID	int	รหัสโมเดล	PK	
MODEL_NAME	varchar (50)	ชื่อโมเดล		
CREATE_DATE	date	วันที่สร้าง		

ตารางที่ 3.11 แสดงข้อมูล Entity ของทรี (Tree)

Attribute	Type	Detail	Key	Reference
TREE_ID	int	รหัสทรี	PK	
MODEL_ID	varchar (50)	รหัสโมเดล	PK	MODEL
PARENT_NODE	varchar (50)	โหนดแม่		
CHILD_NODE	varchar (50)	โหนดลูก		
CURRENT_NODE	varchar (50)	โหนดปัจจุบัน		
BRANCH	varchar (50)	ค่าของสาขา		
CREATE_DATE	date	วันที่สร้าง		

ตารางที่ 3.12 แสดงข้อมูล Entity ของกฎ (Rule)

Attribute	Type	Detail	Key	Reference
RULE_ID	int	รหัสกฎ	PK	
MODEL_ID	varchar (50)	ชื่อโมเดล	PK	MODEL
RULE_VALUE	varchar (50)	กฎ		
CREATE_DATE	date	วันที่สร้าง		

บทที่ 4
การประยุกต์ใช้ดาต้าไมนิ่งโดยใช้เทคนิคต้นไม้ตัดสินใจ

ในการประยุกต์ใช้ดาต้าไมนิ่งโดยใช้เทคนิคต้นไม้ตัดสินในนี้ ได้นำข้อมูลของนักเรียนมาทำการวิเคราะห์เพื่อช่วยตัดสินใจในการเลือกสาขาของนักเรียน และเพื่อเป็นประโยชน์ในการวางแผนการเรียนต่อไป

4.1 การกำหนดวัตถุประสงค์ของงาน (Business Objective Determination)

เพื่อการศึกษาเทคนิคของต้นไม้ตัดสินใจเพื่อช่วยตัดสินใจในการเลือกสาขาของนักเรียน และสามารถนำไปใช้ในการวางแผนการเรียนในอนาคตหรือระดับสูงต่อไป ซึ่งจะเป็นประโยชน์สำหรับนักเรียน ผู้ปกครอง ในการที่จะเลือกสาขาการเรียนเพื่อเข้าศึกษาต่อ และสำหรับอาจารย์ในการแนะแนวด้านการเรียนของนักเรียน และผู้บริหารสถานศึกษา เพื่อนำข้อมูลที่ได้ไปใช้ในการจัดการภายในสถานศึกษาต่อไป

4.2 การเตรียมข้อมูล (Data Preparation)

ข้อมูลที่นำมาใช้ในการวิเคราะห์นี้ เป็นข้อมูลของนักเรียนโรงเรียนพิบูลวิทยาลัย จังหวัดลพบุรี ที่เก็บรวบรวมไว้ตั้งแต่ปีการศึกษา 2549 จนถึง 2551 ทั้งส่วนที่เป็นประวัติ ผลการเรียนก่อนเลือกสาขา และผลการเรียนเมื่อจบการศึกษาตามสาขาที่เลือกเรียน โดยทำการเลือกแอตทริบิวต์ในการวิเคราะห์เพื่อช่วยตัดสินใจในการเลือกสาขาของนักเรียน ดังตารางที่ 4.1

ตารางที่ 4.1 แสดงรายละเอียดของแต่ละ Entity

Table Name	Table Description
Students	เก็บประวัตินักเรียน
ResultStudy	เก็บข้อมูลผลการเรียน

สามารถแสดงรายละเอียดในพจนานุกรมข้อมูลได้ ดังนี้

ตารางที่ 4.2 แสดงดาต้าดิกชันนารีของ Entity Students

Attribute	Type	Detail	Key	Reference
ID	int	รหัส	PK	
StudentID	varchar (50)	รหัสนักเรียน		

ตารางที่ 4.2 (ต่อ)

Attribute	Type	Detail	Key	Reference
sGPA_MA	float	คะแนนกลุ่มวิชาคณิตศาสตร์		
sGPA_SCI	float	คะแนนกลุ่มวิชาวิทยาศาสตร์		
sGPA_ENG	float	คะแนนกลุ่มวิชาอังกฤษ		
sGPA_SO	float	คะแนนกลุ่มวิชาสังคม		
sGPA_TH	float	คะแนนกลุ่มวิชาภาษาไทย		
sGPAX	float	คะแนนตลอดหลักสูตร		
sGender	float	เพศ		
sMor	float	ระดับชั้นมัธยม1-3		

ตารางที่ 4.3 แสดงดาต้าดิกชันนารีของ Entity ResultStudy

Attribute	Type	Detail	Key	Reference
StudentID	varchar (50)	รหัสนักเรียน	PK	
sGPA_MA	float	คะแนนเฉลี่ยกลุ่มวิชาคณิตศาสตร์		
sGPA_SCI	float	คะแนนเฉลี่ยกลุ่มวิชาวิทยาศาสตร์		
sGPA_ENG	float	คะแนนเฉลี่ยกลุ่มวิชาอังกฤษ		
sGPA_SO	float	คะแนนเฉลี่ยกลุ่มวิชาสังคม		
sGPA_TH	float	คะแนนเฉลี่ยกลุ่มวิชาภาษาไทย		
sGPAX	float	คะแนนเฉลี่ยตลอดหลักสูตร		
sGender	float	เพศ		
sMajor	float	สาขาการเรียน		
sTarget	float	ผลลัพธ์เป้าหมาย		

ตารางที่ 4.4 แสดงดาต้าดิกชันนารีของ Entity Major

Attribute	Type	Detail	Key	Reference
sMajor	varchar(3)	รหัสสาขาการเรียน	PK	
Major_Name	varchar (50)	ชื่อสาขาการเรียน		

ตารางที่ 4.5 แสดงการกำหนดรหัสสาขาการเรียน

ID	MAJOR
01	วิทย์-คณิต
02	อังกฤษ-คณิต
03	อังกฤษ-จีน
04	อังกฤษ-ฝรั่งเศส
05	ไทย-สังคม

ตารางที่ 4.6 แสดงการแบ่งคลาสของคะแนนผลการเรียนเฉลี่ย

GPAX,GPA	MEAN	CLASS(ช่วงคะแนน)
EXCELLENT	ดีเยี่ยม	3.50-4.00
BEST	ดีมาก	3.00-3.49
AVERAGE	ปานกลาง	2.50-2.99
FAIRLY	พอใช้	2.00-2.49
LOW	ต่ำ	ต่ำกว่า 2.00

ตารางที่ 4.7 แสดงการแบ่งคลาสของผลการเรียนเฉลี่ยแต่ละกลุ่มวิชา

คะแนนเฉลี่ย กลุ่มวิชา	CLASS (ช่วงของผลการเรียน)
EXCELLENT	3.50-4.00
BEST	3.00-3.49
AVERAGE	2.50-2.99
FAIRLY	2.00-2.49
LOW	ต่ำกว่า 2.00

ตารางที่ 4.8 แสดงตัวอย่างข้อมูลนักเรียน

STUDENT_ID	GENDER	MAJOR_ID	THAI	MATH	SCI	SOC	LANG	TECHNO	ART	GPAX	GPA
1001	F	01	EXT.	G	G	F	F	F	F	G	G
1002	F	02	G	G	G	F	F	G	G	F	G
1003	M	04	EXT.	L	F	G	G	G	F	G	G
1004	F	05	G	L	F	G	F	G	F	F	F
1005	M	05	EXT.	F	G	G	F	F	G	G	G

4.3 การทำดาต้าไมนิ่ง (Data Mining)

เทคนิคที่จะนำมาใช้ในการวิเคราะห์ข้อมูลคือ Decision Tree มาแบ่งกลุ่มข้อมูลที่มีอยู่ เพื่อจะทำนายข้อมูล โดยมีอัลกอริทึมที่นำมาใช้ในการสร้าง Decision Tree คือ อัลกอริทึม ID3 และ C4.5 ที่ได้กล่าวไว้ในบทที่ 2

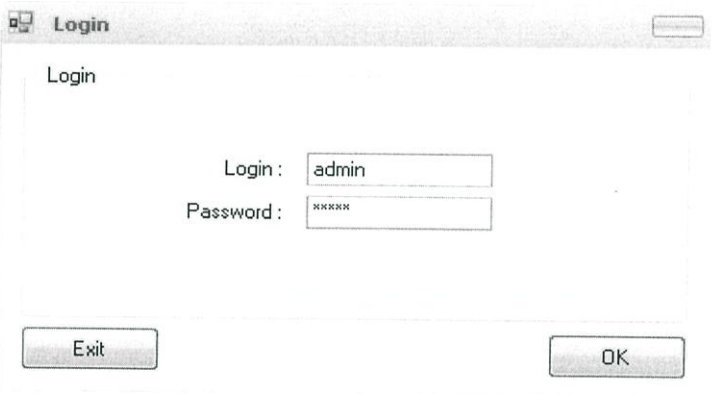
4.4 การพัฒนาระบบงาน

ระบบงานที่พัฒนาขึ้นเป็นส่วนที่ใช้ประมวลผลในการวิเคราะห์รายงาน โดยต้องมีการจัดเตรียมข้อมูลในการสร้างชุดข้อมูลตัวอย่าง (Training Data Set) ประกอบด้วยแอตทริบิวท์ที่เป็น Target Attribute ซึ่งเป็นแอตทริบิวท์ที่ต้องการแบ่งกลุ่มข้อมูลหรือจุดมุ่งหมายที่เราต้องการ และเลือกแอตทริบิวท์ที่เกี่ยวข้องในการสร้างเงื่อนไขหรือสร้างกฎ การทำงานของระบบมีขั้นตอนการทำงาน ดังนี้

- การทำงานของระบบจะมีฟังก์ชั่นการทำงาน 3 ส่วนหลัก คือ
  1. Login เป็นส่วนที่ใช้ในการตรวจสอบผู้ใช้งานระบบ
  2. Models เป็นส่วนที่ใช้ในการจัดเตรียมข้อมูลและสร้างโมเดล Tree
  3. Test Model เป็นส่วนของการทดสอบโมเดลหรือจัดแบ่งประเภทข้อมูลที่ได้จากการสร้างโมเดล Tree

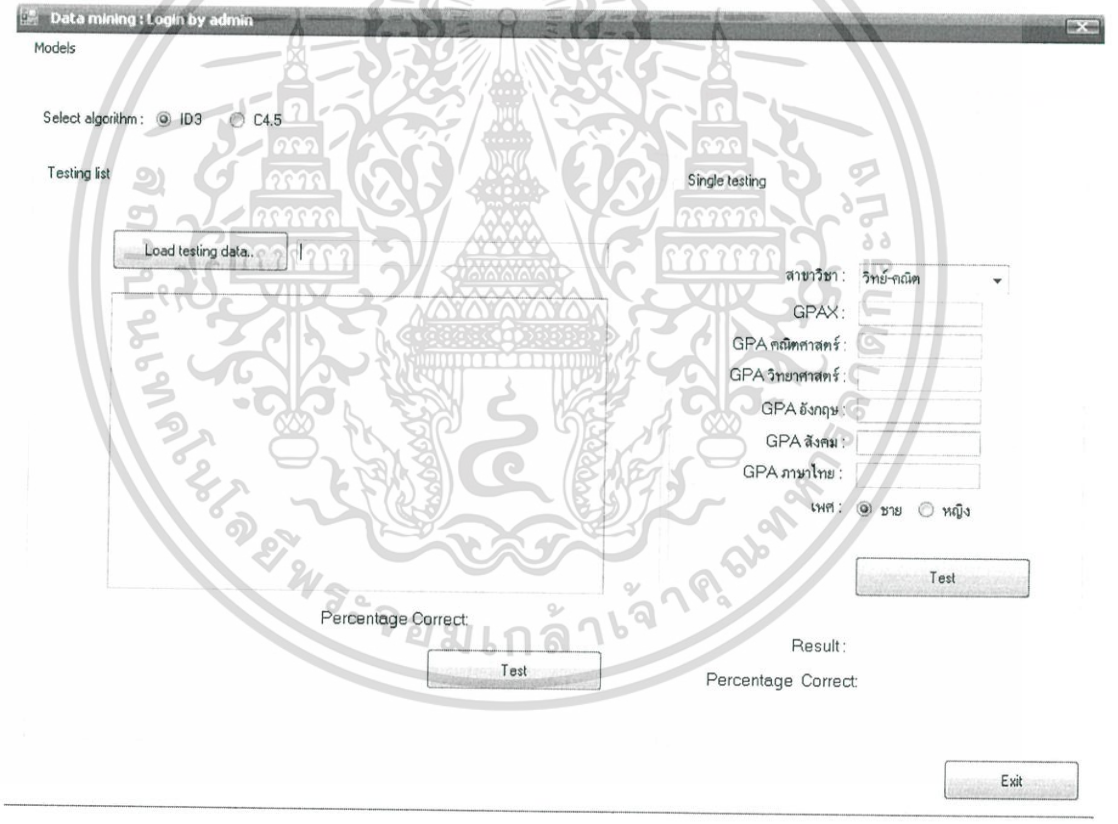
  - ส่วนแสดงการตรวจสอบผู้ใช้งานระบบ (Login)

    ผู้ใช้ทำการตรวจสอบสิทธิ์ในการเข้าใช้งานระบบ โดยจะต้องมีการกรอก Username, Password ก่อนเข้าใช้งานระบบ ดังรูปที่ 4.1



รูปที่ 4.1 หน้อจอแสดงส่วน Login

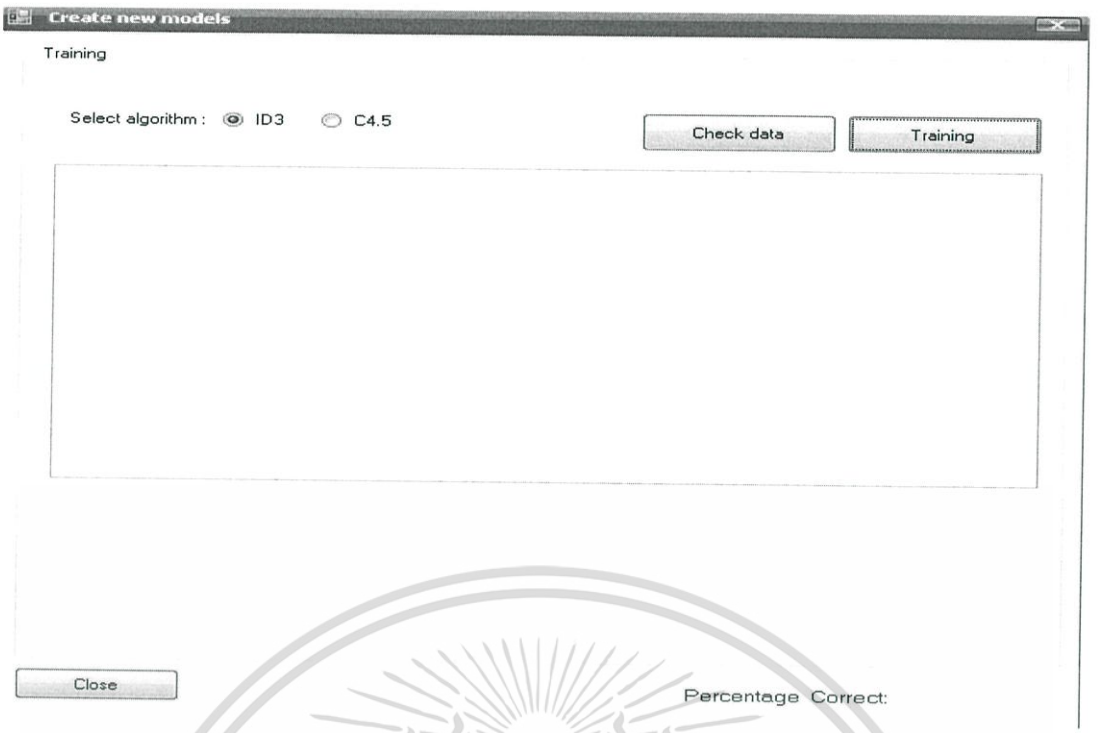
ในกรณีที่ผู้ใช้งานระบบเป็น Admin จะแสดงหน้าจอ Models ซึ่งเป็นส่วนที่ใช้ในการจัดเตรียมข้อมูลและสร้างโมเดล Tree ดังรูปที่ 4.2



รูปที่ 4.2 หน้อจอแสดงส่วน Models

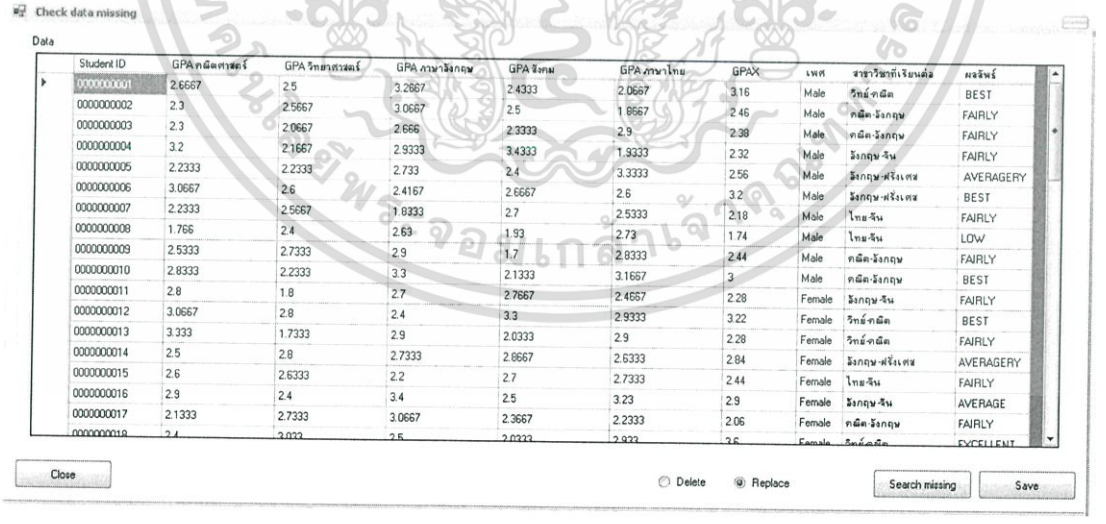
- ส่วนแสดงการจัดเตรียมข้อมูลและสร้างโมเดล Tree ในระบบงาน

ผู้ใช้จะทำการจัดเตรียมข้อมูลและสร้างโมเดล Tree โดยกดปุ่ม "Models" เลือก "Create new models" ดังรูปที่ 4.3



รูปที่ 4.3 หน้าจอแสดงส่วน Training

ผู้ใช้ทำการตรวจสอบข้อมูลในระบบโดยกดปุ่ม "Check Data" กรณีที่พบว่ามีข้อมูล Missing Values จะต้องทำการจัดการกับข้อมูล Missing Values ก่อน โดยสามารถทำการลบ record ทิ้งไป หรือแทนค่าได้ตามคุณสมบัติของ Data Type ของแอตทริบิวท์นั้นๆ ดังรูปที่ 4.4

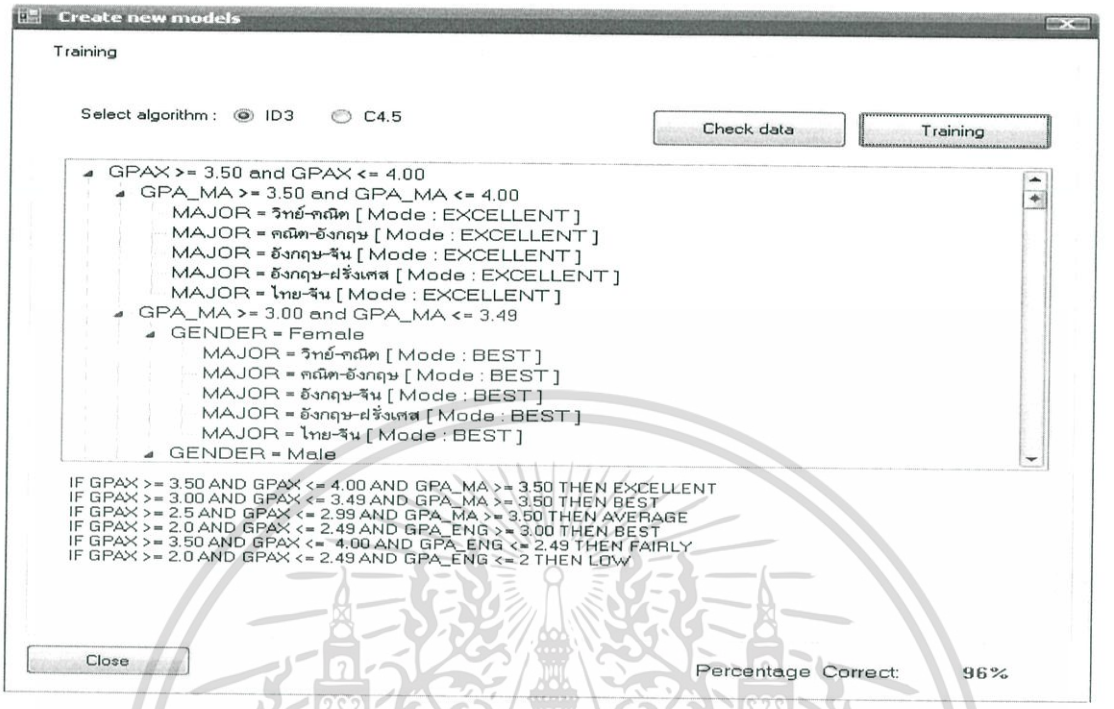


Student ID	GPA คณิตศาสตร์	GPA วิทยาศาสตร์	GPA ภาษาอังกฤษ	GPA สังคม	GPA ภาษาไทย	GPAX	เพศ	สาขาวิชาที่เรียนต่อ	ผลลัพธ์
0000000001	2.6667	2.5	3.2667	2.4333	2.0667	3.16	Male	วิทย์-คณิต	BEST
0000000002	2.3	2.5667	3.0667	2.5	1.9667	2.46	Male	คณิต-อังกฤษ	FAIRLY
0000000003	2.3	2.0667	2.666	2.3333	2.9	2.38	Male	คณิต-อังกฤษ	FAIRLY
0000000004	3.2	2.1667	2.9333	3.4333	1.9333	2.32	Male	อังกฤษ-จีน	FAIRLY
0000000005	2.2333	2.2333	2.733	2.4	3.3333	2.56	Male	อังกฤษ-ฝรั่งเศส	AVERAGERY
0000000006	3.0667	2.6	2.4167	2.6667	2.6	3.2	Male	อังกฤษ-ฝรั่งเศส	BEST
0000000007	2.2333	2.5667	1.8333	2.7	2.5333	2.18	Male	ไทย-จีน	FAIRLY
0000000008	1.766	2.4	2.63	1.93	2.73	1.74	Male	ไทย-จีน	LOW
0000000009	2.5333	2.7333	2.9	1.7	2.8333	2.44	Male	คณิต-อังกฤษ	FAIRLY
0000000010	2.8333	2.2333	3.3	2.1333	3.1667	3	Male	คณิต-อังกฤษ	BEST
0000000011	2.8	1.8	2.7	2.7667	2.4667	2.28	Female	อังกฤษ-จีน	FAIRLY
0000000012	3.0667	2.8	2.4	3.3	2.9333	3.22	Female	วิทย์-คณิต	BEST
0000000013	3.333	1.7333	2.9	2.0333	2.9	2.28	Female	วิทย์-คณิต	FAIRLY
0000000014	2.5	2.8	2.7333	2.8667	2.6333	2.84	Female	อังกฤษ-ฝรั่งเศส	AVERAGERY
0000000015	2.6	2.6333	2.2	2.7	2.7333	2.44	Female	ไทย-จีน	FAIRLY
0000000016	2.9	2.4	3.4	2.5	3.23	2.9	Female	อังกฤษ-จีน	AVERAGE
0000000017	2.1333	2.7333	3.0667	2.3667	2.2333	2.06	Female	คณิต-อังกฤษ	FAIRLY
0000000018	2.4	3.033	2.5	2.0333	2.933	2.5	Female	วิทย์-คณิต	EXCELLENT

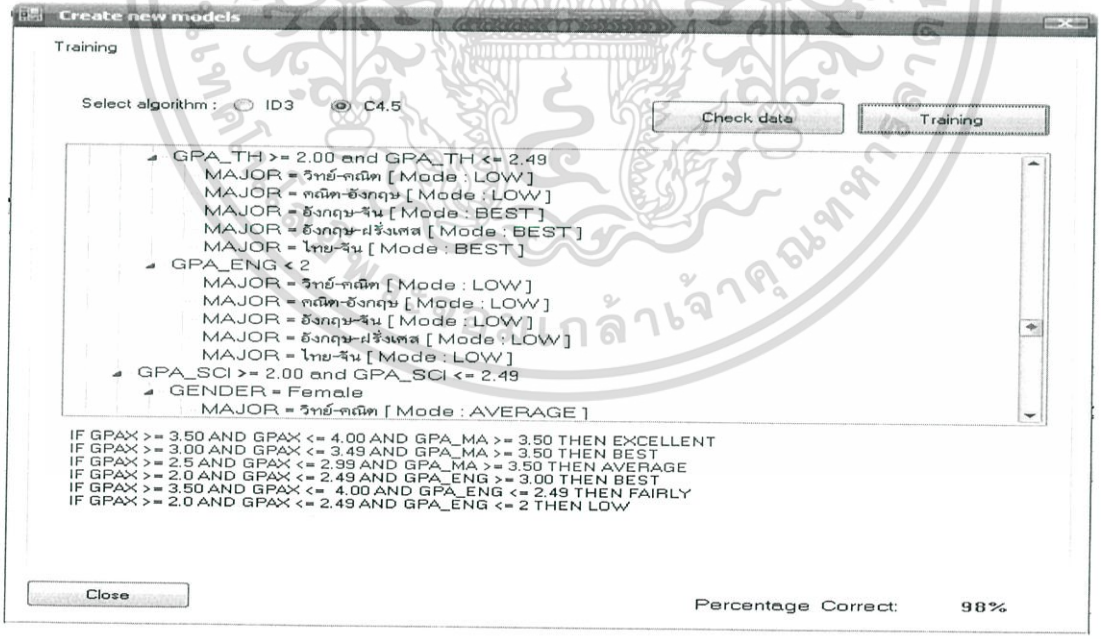
รูปที่ 4.4 หน้าจอแสดงส่วน Missing Values

เมื่อทำการตรวจสอบข้อมูลในระบบเป็นที่เรียบร้อยแล้ว จะเป็นขั้นตอนของการสร้างโมเดล Tree จากข้อมูลเหล่านั้น โดยผู้ใช้จะทำการเลือกอัลกอริทึม Classifier Tree (ID3,C4.5)

และกดปุ่ม “Training” ระบบจะทำการสร้างโมเดล และแสดงผลในรูปแบบของต้นไม้ตัดสินใจ ดังรูปที่ 4.5 และรูปที่ 4.6

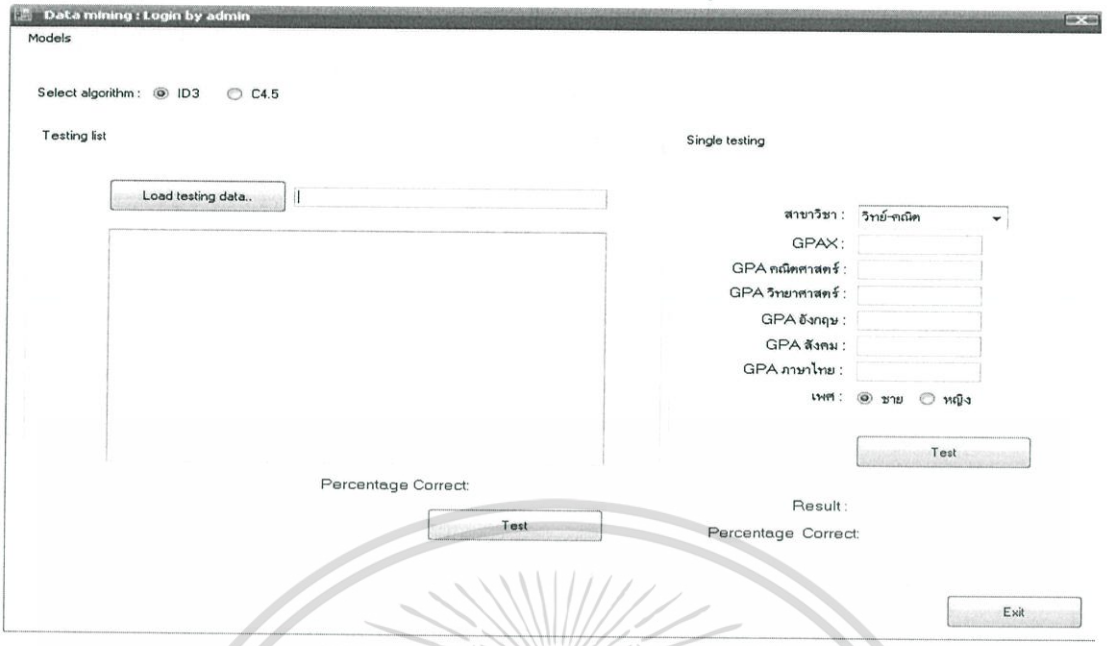


รูปที่ 4.5 หน้าจอแสดงการสร้างโมเดล Tree ในกรณีเลือกอัลกอริทึม ID3



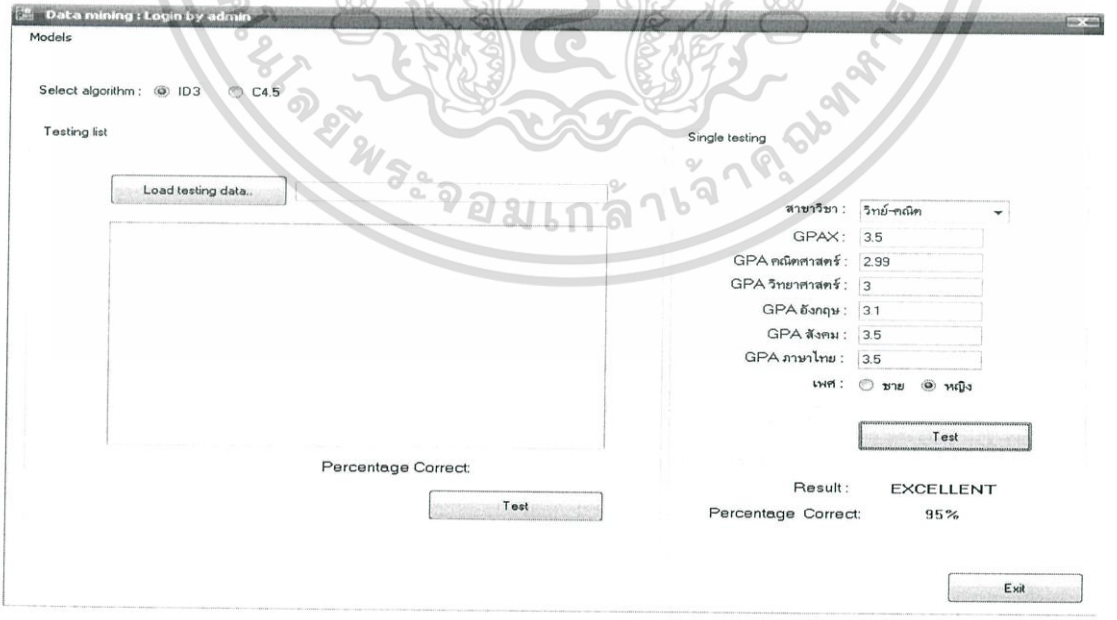
รูปที่ 4.6 หน้าจอแสดงการสร้างโมเดล Tree ในกรณีเลือกอัลกอริทึม C4.5

- ส่วนแสดงการนำโมเดลที่ได้ไปทดสอบ (Test) ดังรูปที่ 4.7

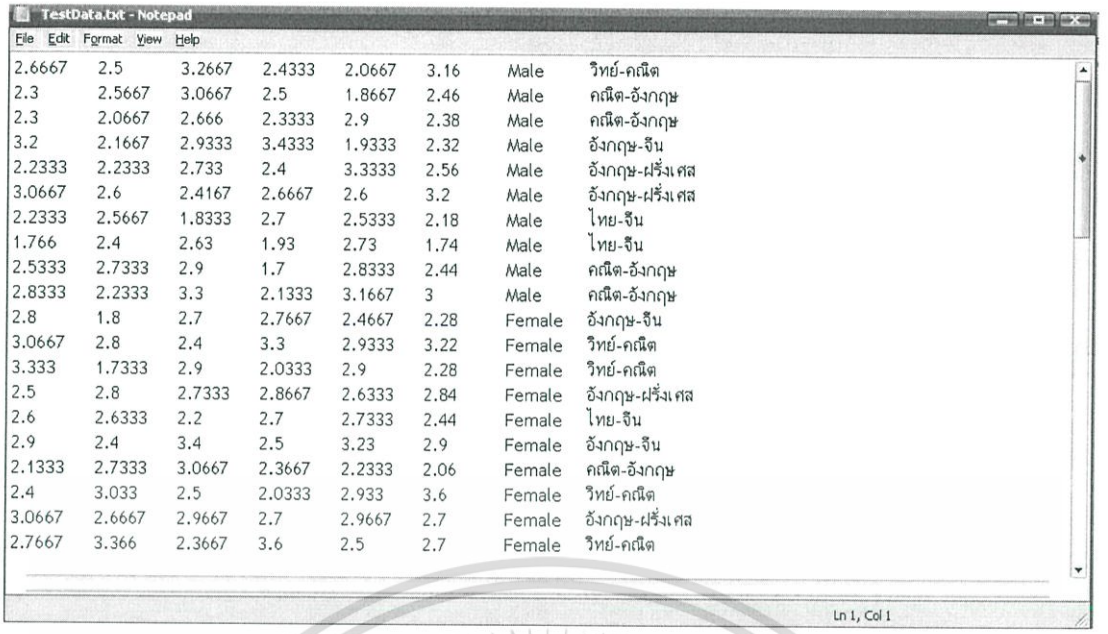


รูปที่ 4.7 หน้าจอแสดงการทดสอบโมเดลที่ได้จากการสร้างโมเดล Tree

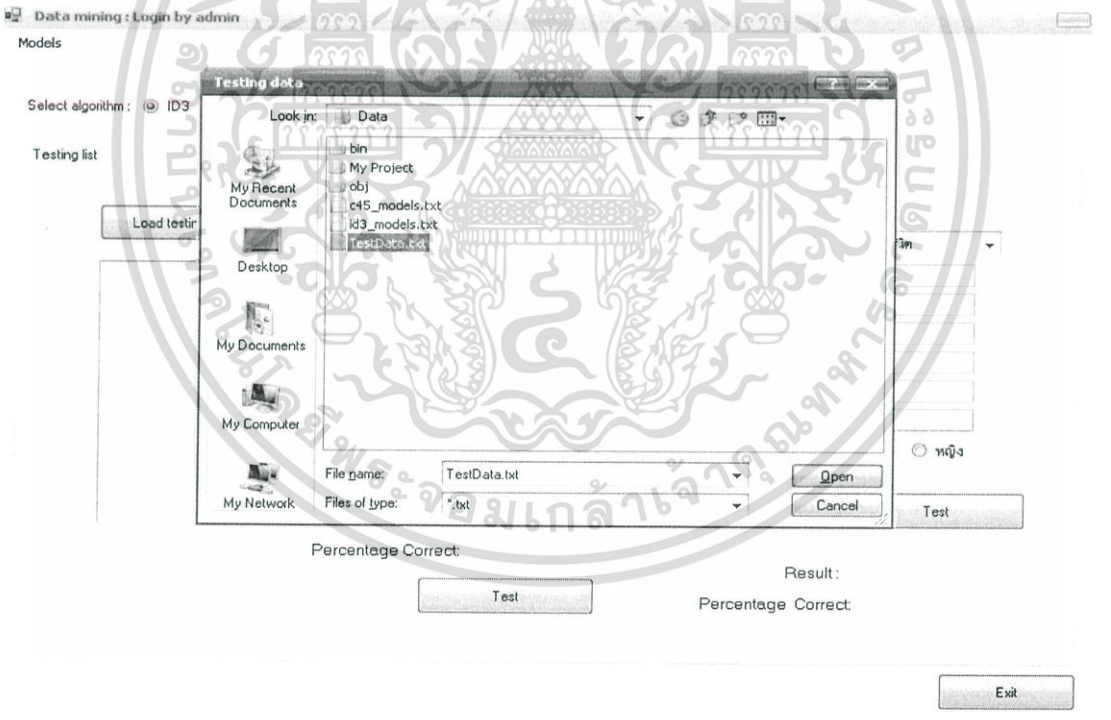
หลังจากที่ได้สร้างโมเดลแล้ว จะเป็นการนำโมเดลที่ได้ไปทำการทดสอบโมเดลหรือจัดแบ่งประเภทข้อมูลที่ได้จากการสร้างโมเดลต้นไม้ โดยการใช้ข้อมูลทดสอบ (Test Data) ซึ่งเป็นคนละชุดกับที่ใช้ในการสร้างชุดข้อมูลตัวอย่าง (Training Data Set) และสามารถทำนายข้อมูลที่ผู้ใช้ทำการกรอกข้อมูลใหม่เข้าไปโดยใช้โมเดลที่เลือก



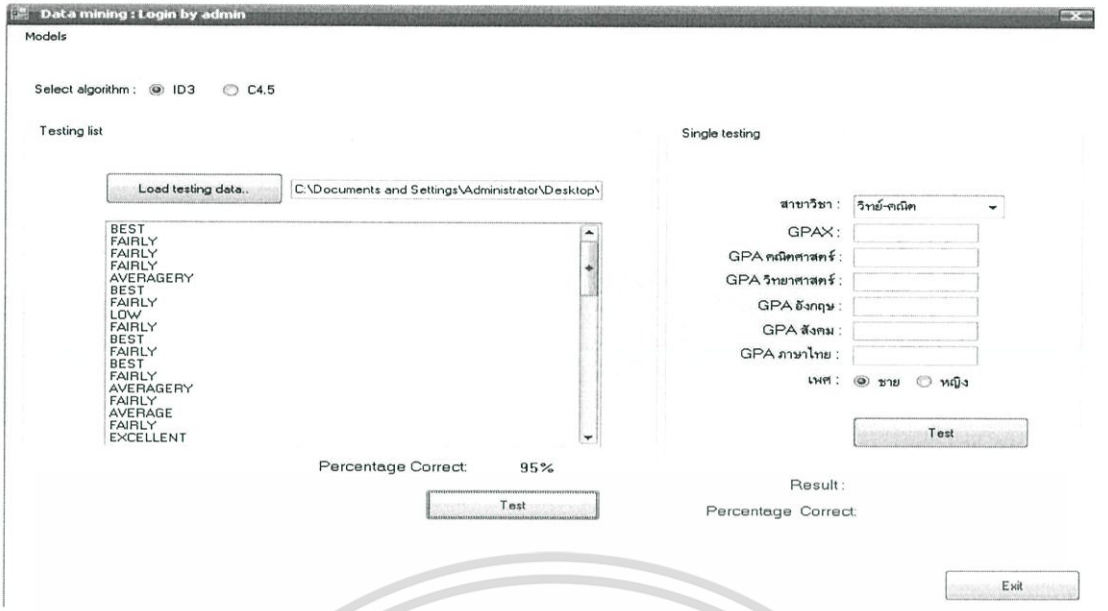
รูปที่ 4.8 หน้าจอแสดงการทดสอบโมเดลที่ได้จากการสร้างโมเดล Tree กรณีทดสอบเป็นรายบุคคล



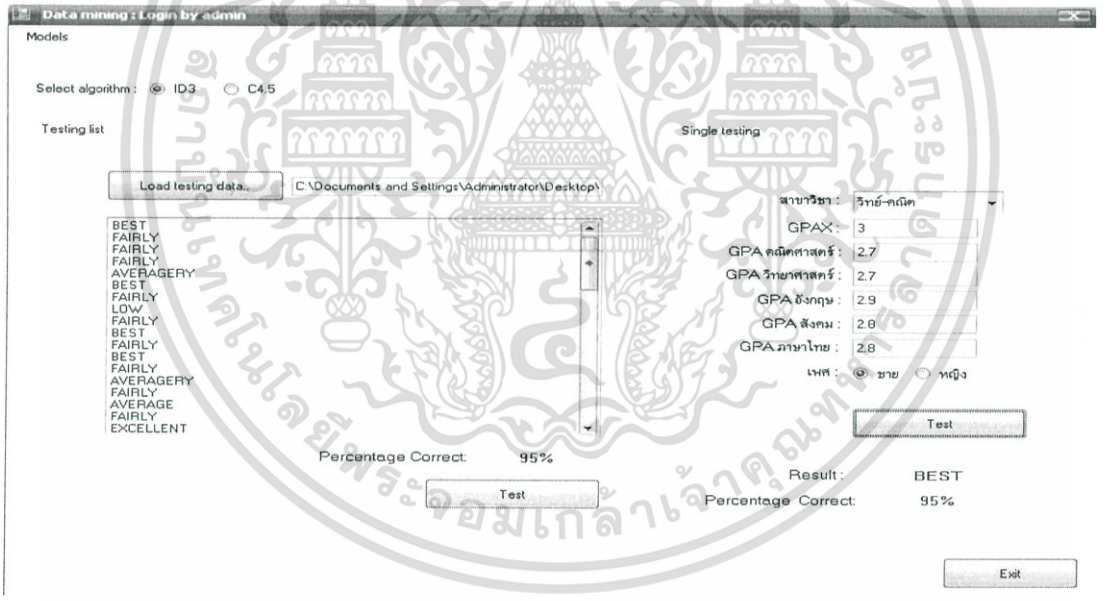
รูปที่ 4.9 หน้าจอแสดงข้อมูลกรณีทดสอบเป็นกลุ่ม



รูปที่ 4.10 หน้าจอแสดงการเลือกข้อมูลที่จะนำมาทดสอบทดสอบ

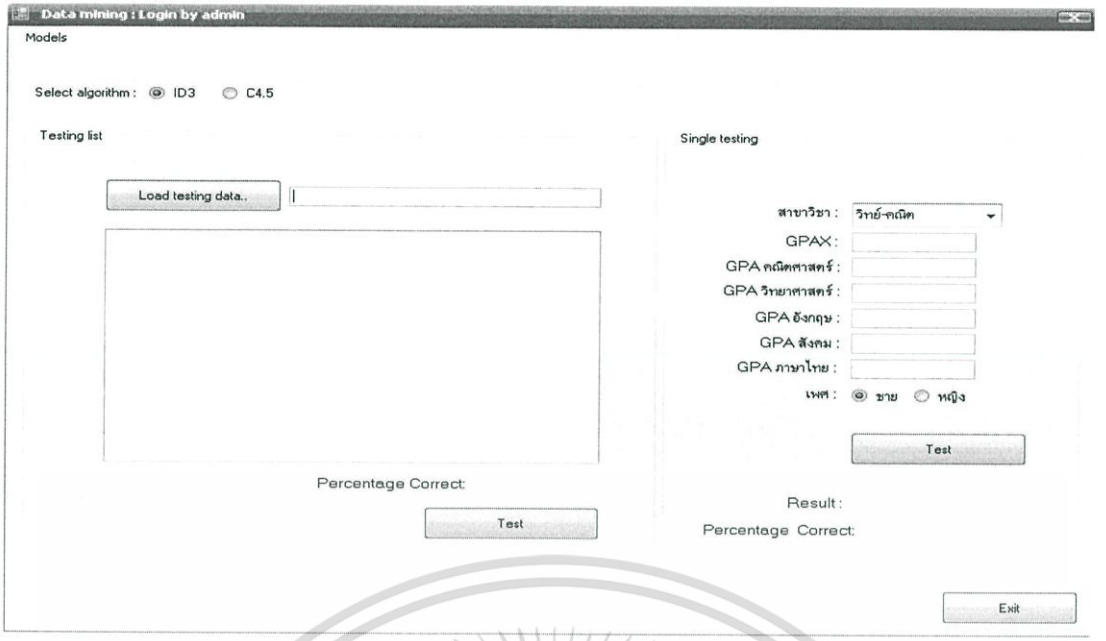


รูปที่ 4.11 หน้าจอแสดงการทดสอบโมเดลที่ได้จากการสร้างโมเดล Tree ทดสอบเป็นกลุ่ม



รูปที่ 4.12 หน้าจอแสดงการทำนายแนวโน้มของผลการเรียนโดยระบุสาขาที่เลือก

ในกรณีที่ผู้ใช้งานระบบเป็น User จะแสดงหน้อจอการนำโมเดลที่ได้ไปทดสอบ (Test) ดังรูปที่ 4.7 โดยสามารถใช้งานในส่วนนี้ได้เหมือนกับในส่วน Admin ซึ่งสามารถแสดงได้ดังรูปข้างบนในส่วนของการทดสอบโมเดลที่ได้จากการสร้างโมเดล Tree



รูปที่ 4.13 หน้าจอแสดงการทดสอบโมเดลที่ได้จากการสร้างโมเดล Tree

4.5 การวิเคราะห์ผลลัพธ์ (Analysis of Results)

ผลลัพธ์ที่ได้จากการสร้าง Decision Tree เพื่อช่วยตัดสินใจในการเลือกสาขาของนักเรียนนั้น ได้นำข้อมูลของนักเรียน โรงเรียนพิบูลวิทยาลัย จังหวัดลพบุรี ตั้งแต่ปีการศึกษา 2549 จนถึง 2551 มาใช้ในประกอบกับการวิเคราะห์ซึ่งผลลัพธ์ที่ได้ เช่น

ผลลัพธ์ที่ได้จากรูปที่ 4.6 สามารถวิเคราะห์ผลลัพธ์ได้ดังนี้ คือ

1. นักเรียนที่มีผลการเรียนเฉลี่ยน้อยกว่า 4.00 และมากกว่าหรือเท่ากับ 3.50 และเป็นเพศหญิง และระดับผลการเรียนกลุ่มวิชาวิทยาศาสตร์ EXCELLENT, กลุ่มวิชาคณิตศาสตร์ EXCELLENT , กลุ่มวิชาอังกฤษ EXCELLENT แล้วเลือกเรียนสาขาวิทย์-คณิต ผลการเรียนเฉลี่ยตลอดหลักสูตรเมื่อสำเร็จการศึกษาคาดว่าจะอยู่ในระดับ EXCELLENT

2. นักเรียนที่มีผลการเรียนเฉลี่ยน้อยกว่า 3.49 และมากกว่าหรือเท่ากับ 3.00และเป็นเพศหญิง และระดับผลการเรียนกลุ่มวิชาวิทยาศาสตร์ BEST, กลุ่มวิชาคณิตศาสตร์ BEST, กลุ่มวิชาอังกฤษ BEST แล้วเลือกเรียนสาขาวิทย์-คณิต ผลการเรียนเฉลี่ยตลอดหลักสูตรเมื่อสำเร็จการศึกษาคาดว่าจะอยู่ในระดับ BEST

นักเรียนที่มีผลการเรียนเฉลี่ยน้อยกว่า 3.49 และมากกว่าหรือเท่ากับ 3.00 และเป็นเพศชาย และระดับผลการเรียนกลุ่มวิชาวิทยาศาสตร์ AVERAGE, กลุ่มวิชาคณิตศาสตร์ AVERAGE, กลุ่มวิชาอังกฤษ BEST

ถ้าเลือกเรียนสาขาวิทย์-คณิต ผลการเรียนเฉลี่ยตลอดหลักสูตรเมื่อสำเร็จการศึกษาคาดว่าจะอยู่ในระดับ AVERAGE

ถ้าเลือกเรียนสาขาอังกฤษ-คณิต ผลการเรียนเฉลี่ยตลอดหลักสูตรเมื่อสำเร็จการศึกษาคาดว่าจะอยู่ในระดับ AVERAGE

ถ้าเลือกเรียนสาขาอังกฤษ-ฝรั่งเศส ผลการเรียนเฉลี่ยตลอดหลักสูตรเมื่อสำเร็จการศึกษาคาดว่าจะอยู่ในระดับ AVERAGE

3. นักเรียนที่มีผลการเรียนเฉลี่ยน้อยกว่า 2.88 และมากกว่าหรือเท่ากับ 2.31 และเป็นเพศชายและระดับผลการเรียนกลุ่มวิชาวิทยาศาสตร์ LOW, กลุ่มวิชาคณิตศาสตร์ LOW, กลุ่มวิชาอังกฤษ FAIRLY แล้วเลือกเรียนสาขาไทย-สังคม ผลการเรียนเฉลี่ยตลอดหลักสูตรเมื่อสำเร็จการศึกษาคาดว่าจะอยู่ในระดับ FAIRLY

ผลลัพธ์ที่ได้จากรูปที่ 4.5 และ 4.6 สามารถวิเคราะห์ผลลัพธ์ได้ดังนี้ คือ อัลกอริทึม C4.5 มีเปอร์เซนต์ความถูกต้องในการสร้างโมเดล Tree เท่ากับ 98% ซึ่งมีค่าความถูกต้องมากกว่าการสร้างโมเดล Tree ด้วยอัลกอริทึม ID3 ซึ่งมีเปอร์เซนต์ความถูกต้องในการสร้างโมเดล Tree เท่ากับ 96% เนื่องจากอัลกอริทึม C4.5 ได้มีการพัฒนามาจากอัลกอริทึม ID3 เพื่อเพิ่มประสิทธิภาพในการทำนายผลให้มีความถูกต้องมากยิ่งขึ้น โดยมีเทคนิคในการแก้ปัญหาที่เกิดจากอัลกอริทึม ID3 เช่น แก้ปัญหาOver-Fitting ในการสร้างต้นไม้ตัดสินใจ เป็นต้น

บทที่ 5
สรุปผลการศึกษา และข้อเสนอแนะ

5.1 สรุปผลการพัฒนาระบบงาน

การพัฒนาระบบนี้ มีวัตถุประสงค์เพื่อศึกษาเทคนิคของต้นไม้ตัดสินใจเพื่อช่วยตัดสินใจในการเลือกสาขาของนักเรียน โดยทำการสร้างโมเดลด้วยอัลกอริทึม ID3 และ C4.5 และนำโมเดลที่ได้มาใช้ในการวิเคราะห์ซึ่งผลลัพธ์ที่ได้จะเป็นประโยชน์ในการช่วยตัดสินใจในการเลือกสาขาของนักเรียน และสามารถนำไปปรับปรุงแผนทางการศึกษาภายในสถานศึกษาให้มีคุณภาพที่ดีขึ้น

จากข้อมูลที่นำมาใช้ในระบบ จะเป็นข้อมูลที่ได้จัดเตรียมข้อมูลให้อยู่ในรูปแบบที่เหมาะสมและมีความถูกต้องก่อนนำมาใช้งาน โดยแต่ละแอตทริบิวต์มีการแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสม และจัดการข้อมูลที่ไม่ถูกต้องก่อนนำมาใช้งาน โดยระบบสามารถรองรับข้อมูลที่เป็นทั้ง Categorical และ Quantitative หลังจากนั้นจะทำการสร้างโมเดลด้วยอัลกอริทึม ID3 และ C4.5 ซึ่งจากการแสดงผลที่ได้จากรูปที่ 4.5 และ 4.6 สามารถวิเคราะห์ผลลัพธ์ได้ดังนี้ คือ อัลกอริทึม C4.5 มีเปอร์เซนต์ความถูกต้องในการสร้างโมเดล Tree เท่ากับ 98% ซึ่งมีค่าความถูกต้องมากกว่าการสร้างโมเดล Tree ด้วยอัลกอริทึม ID3 ซึ่งมีเปอร์เซนต์ความถูกต้องในการสร้างโมเดล Tree เท่ากับ 96% เพราะอัลกอริทึม C4.5 มีเทคนิคในการแก้ปัญหาที่เกิดจากอัลกอริทึม ID3 เช่น แก้ปัญหาOver-Fitting ในการสร้างต้นไม้ตัดสินใจ โดยใช้เทคนิคการ Pruning ทำให้โมเดล Tree ที่มีความถูกต้องมากยิ่งขึ้น เป็นต้น

5.2 ข้อเสนอแนะ

5.2.1 ข้อมูลที่นำมาใช้ในการ Training ควรเป็นข้อมูลที่ถูกต้อง สมบูรณ์ เพื่อทำให้ประสิทธิภาพในการทำนายผลมีความถูกต้องแม่นยำมากขึ้น และตรงกับความต้องการของผู้ใช้มากที่สุด

5.2.2 สามารถนำไปประยุกต์ใช้ และปรับปรุงในงานที่มีลักษณะคล้ายคลึงกันได้ เช่น การคาดคะเนผลการเรียนในรายวิชาต่าง ๆ โดยอาจเลือกเทคนิคอื่นในการทำดาต้าไมนิ่ง เพื่อให้ได้โมเดลที่มีความเหมาะสมกับข้อมูลที่นำมาวิเคราะห์มากที่สุด

5.2.3 โมเดลที่ใช้ในการพัฒนาระบบนี้ สามารถมีปัจจัยอื่นที่ส่งผลต่อการเลือกสาขาของนักเรียนได้อีก เช่น การเลือกเรียนสาขาตามเพื่อน การเลือกสาขาตามค่านิยมในสังคม

บรรณานุกรม

บุญเสริม กิจสิริกุล. 2546. **โครงการวิจัยร่วมภาครัฐและเอกชน : โครงการย่อยที่ 7 อัลกอริทึมการทำเหมืองข้อมูล.** ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์ มหาวิทยาลัย.

บุญมา เพ่งซ่วน. 2548. **การใช้เทคนิคเหมืองข้อมูลเพื่อพัฒนาระบบสนับสนุนการตัดสินใจด้านการผลิตบัณฑิตระดับปริญญาตรี.** วิทยานิพนธ์วิทยาศาสตร์มหาบัณฑิต สาขาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร.

พงษ์พันธ์ ศิวิลัย. 2552. **SQL SERVER 2008 ฉบับสมบูรณ์.** กรุงเทพ: ซีเอ็ดยูเคชั่น.

Cabena.et al. 1998. **Discovering Data Mining.** New Jersy : Prentice Hall.

Han.et al. 2001. **Data Mining Concepts and Techniques.** SanFrancisco : Morgan Kaufmann.

Leo Breiman.et al. 1984. **Classification and regression trees.**

Occam, Razor. **Decision Trees & Data Mining.** [Online]. Available: http://www.decisiontrees.net/node/27.

Simon, Colton. 2004. **Decision Tree.** [Online]. Available: http://www.doc.ic.ac.uk/~sgc/teaching/v231/lecture11.html.

Winston P. 1992. **C4.5 Tutorial.** [Online]. Available: http://www2.cs.uregina.ca/~hamilton/courses/831/notes/ml/dtrees/c4.5/tutorial.html.

ประวัติผู้เขียน

ชื่อผู้เขียน	นางสาวตะวัน ระวิงทอง
สถานที่เกิด	กรุงเทพมหานคร
ประวัติการศึกษา	จบการศึกษาระดับปริญญาตรี วิทยาศาสตรบัณฑิต สาขาเทคโนโลยีคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคล ธัญบุรี