การเปรียบเทียบประสิทธิภาพในการจำแนกกลุ่มเมื่อข้อมูลมีค่านอกเกณฑ์ในการทำเหมืองข้อมูล

PERFORMANCE COMPARISON OF DATA MINING'S CLASSIFICATION METHODS ON DATA SET WITH OUTLIERS

พนิดา สมบัติมาก

ภัสสร จันทร์หอม

ศุภกร รัศมี

โอฬาร รุ่งมณีธรรมคุณ

ปัญหาพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรบัณฑิต (สถิติประยุกต์)

ภาควิชาสถิติ คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ปีการศึกษา 2560

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

PERFORMANCE COMPARISON OF DATA MINING'S CLASSIFICATION METHODS ON DATA SET WITH OUTLIERS

PHANIDA SOMBATMAK
PATSORN JANHOM
SUPHAKORN RATSAMEE
ORAN RUNGMANEETHUMMAKUN

A SPECIAL PROBLEM SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR

THE DEGREE OF BACHELOR OF SCIENCE (APPLIED STATISTICS)

DEPARTMENT OF STATISTICS, FACULTY OF SCIENCE

KING MONGKUT'S INTITUTE OF TECHNOLOGY LADKRABANG

ACADEMIC YEAR 2017

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

หัวข้อปัญหาพิเศษ	การเปรียบเทียบประสิทธิภาพในการจำแนกกลุ่มเมื่อข้อมูลมีค่านอกเกณฑ์ในการทำเหมืองข้อมูล Performance Comparison of Data Mining's Classification Methods on Data Set with Outliers		
ชื่อนักศึกษา	นางสาวพนิดา	สมบัติมาก	รหัสนักศึกษา 57051143
	นางสาวภัสสร	จันทร์หอม	รหัสนักศึกษา 57051152
	นายศุภกร	รัศมี	รหัสนักศึกษา 57051177
	นายโอฬาร	รุ่งมณีธรรมคุณ	รหัสนักศึกษา 57051202
ปริญญา	วิทยาศาสตรบัณฑิต (สถิติประยุกต์)		
ภาควิชา	สถิติ		
ปีการศึกษา	2560		
อาจารย์ที่ปรึกษา	รองศาสตราจารย์สายชล สินสมบูรณ์ทอง		

คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง อนุมัติให้ปัญหาพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรบัณฑิต (สถิติประยุกต์) ประจำปีการศึกษา 2560

คณะกรรมการสอบ	ลายมือชื่อ
ดร.พรพิมล ชัยวุฒิศักดิ์ ประธานกรรมการ	
ดร.ยุวดี กล่อมวิเศษ กรรมการ	
รองศาสตราจารย์ สายชล สินสมบูรณ์ทอง กรรมการและอาจารย์ที่ปรึกษา	

ลิขสิทธิ์ของคณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

หัวข้อปัญหาพิเศษ	การเปรียบเทียบประสิทธิภาพในการจำแนกกลุ่มเมื่อข้อมูลมีค่านอกเกณฑ์ในการทำเหมืองข้อมูล		
ชื่อนักศึกษา	นางสาวพนิดา	สมบัติมาก	รหัสนักศึกษา 57051143
	นางสาวภัสสร	จันทร์หอม	รหัสนักศึกษา 57051152
	นายศุภกร	รัศมี	รหัสนักศึกษา 57051177
	นายโอฬาร	รุ่งมณีธรรมคุณ	รหัสนักศึกษา 57051202
ปริญญา	วิทยาศาสตร์บัณฑิต (สถิติประยุกต์)		
ภาควิชา	สถิติ		
ปีการศึกษา	2560		
อาจารย์ที่ปรึกษา	รองศาสตราจารย์สายชล สินสมบูรณ์ทอง		

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพในการจำแนกกลุ่ม 5 วิธี คือ วิธีนาอีฟเบส์ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม และวิธีซัพพอร์ตเวกเตอร์แมชชีน โดยพิจารณาจากค่าความถูกต้อง ค่าความคาดเคลื่อนกำลังสองเฉลี่ยและค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย และเพื่อเปรียบเทียบวิธีการสุ่มตัวอย่างระหว่างโปรแกรม SPSS และ WEKA โดยแบ่งข้อมูลเป็นชุดข้อมูลเรียนรู้ ชุดข้อมูลตรวจสอบความถูกต้อง และชุดข้อมูลทดสอบ ในอัตราส่วน 70, 20 และ 10 ตามลำดับ โดยการค้นคว้าและศึกษาในการหาค่านอกเกณฑ์ได้ข้อมูล 3 ชุด คือโรคมะเร็งเต้านมของรัฐวิสคอนซิน เป็นชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับต่ำ โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เป็นชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับปานกลาง และการชำระเงินด้วยบัตรเครดิตของลูกค้า เป็นชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับสูง โดยใช้เครื่องมือ Highlight Exceptions ในการตรวจจับค่านอกเกณฑ์

จากการเปรียบเทียบข้อมูล โรคมะเร็งเต้านมของรัฐวิสคอนซิน วิธีที่มีประสิทธิภาพสูงสุดคือ วิธีโครงข่ายประสาทเทียม โดยการสุ่มของโปรแกรม SPSS โรคเบาหวานของชาวพิม่า ประเทศอินเดีย วิธีที่มีประสิทธิภาพสูงสุดคือ วิธีเพื่อนบ้านใกล้สุด k ตัว โดยการสุ่มของโปรแกรม SPSS และ WEKA และการชำระเงินด้วยบัตรเครดิตของลูกค้า วิธีที่มีประสิทธิภาพสูงสุดคือ วิธีเพื่อนบ้านใกล้สุด k ตัว โดยการสุ่มของโปรแกรม SPSS และ WEKA ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับปานกลางและสูงให้ผลการจำแนกกลุ่มที่เหมือนกัน ซึ่งแตกต่างจากชุดข้อมูลที่มีค่านอกเกณฑ์ในระดับที่ต่ำ

คำสำคัญ : ค่านอกเกณฑ์ วิธีนาอีฟเบส์ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม วิธีซัพพอร์ตเวกเตอร์แมชชีน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Title	Performance Comparison of Data Mining's Classification Methods on Data Set with Outliers		
Students	PHANIDA	SOMBATMAK	Student ID 57051143
	PATSORN	JANHOM	Student ID 57051152
	SUPHAKORN	RATSAMEE	Student ID 57051177
	ORAN	RUNGMANEETHUMMAKUN	Student ID 57051202
Degree	Bachelor of Science (Applied Statistics)		
Department	Statistics		
Faculty	Science		
University	King Mongkut's Institute of Technology Ladkrabang (KMITL)		
Academic Year	2017		
Advisor	Assoc.Prof.Saichon Sinsomboonthong		

ABSTRACT

The objectives of this study were to evaluate and compare the performances of 5 classification methods: naïve bayes, k-nearest neighbors, decision tree, artificial neural network, and support vector machine and to compare the sampling methods by SPSS and WEKA. The performance measures were prediction accuracy, mean squared error, and mean absolute deviation. In sampling methods comparison, the data set used were a data set on the prevalence of breast cancer in Wisconsin, USA, another data set on the prevalence of diabetes in Pima people, India, and another one on Taiwanese customer's payment through credit card. Each of these data sets were divided into three smaller sets: training, validating, and testing sets at a proportion of 70:20:10. Using Highlight Exceptions add-in to examine outliers

For the prevalence of breast cancer data set, the best classification method was the artificial neural network method in combination with the SPSS sampling method. For both the prevalence of diabetes and payment through credit card data sets, the best classification method was the k-nearest neighbors method in combination with either SPSS or WEKA sampling method. The data sets that had a moderate to high number of outliers favored the same classification method while the data set that had a low number of outliers did not favor the same classification method as those two mentioned above.

Keywords: outlier, naïve Bayes, k-nearest neighbors, decision tree, artificial neural network, support vector machine.

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

กิตติกรรมประกาศ

ปัญหาพิเศษฉบับนี้สำเร็จลุล่วงไปได้ด้วยดี มีความละเอียดและความถูกต้องในเนื้อหา เนื่องด้วยได้รับความอนุเคราะห์จาก รศ.สายชล สินสมบูรณ์ทอง (อาจารย์ที่ปรึกษา) ที่ได้ให้ความรู้ ให้คำปรึกษา ให้ความช่วยเหลือ และตรวจแก้ไขในการทำปัญหาพิเศษจนสำเร็จลุล่วงไปได้ด้วยดี

ขอขอบคุณ ดร.พรพิมล ชัยวุฒิศักดิ์ และ ดร.ยุวดี กล่อมวิเศษ ที่เป็นอาจารย์คณะกรรมการที่ให้คำแนะนำเกี่ยวกับปัญหาพิเศษฉบับนี้ทั้งหมด ทั้งแนวคิด วิธีการ และข้อมูลทุกอย่างที่เป็นประโยชน์ในการวิเคราะห์วิธีการจำแนกกลุ่ม

ขอขอบคุณอาจารย์ภาคสถิติทุกท่าน ที่ได้ประสิทธิ์ประสาทวิชาความรู้ พร้อมทั้งคำแนะนำและช่วยเหลือในเรื่องต่างๆตลอดมา

สุดท้ายนี้ขอขอบคุณครอบครัวของผู้จัดทำ ซึ่งสนับสนุนในด้านกำลังทรัพย์ และให้กำลังใจเสมอมา ขอบคุณเพื่อนๆทุกคนที่ให้คำปรึกษา ช่วยเหลือในการทำงานมาโดยตลอด จนปัญหาพิเศษฉบับนี้สำเร็จลุล่วงด้วยดี

พนิดา สมบัติมาก
ภัสสร จันทร์หอม
ศุภกร รัศมี
โอฬาร รุ่งมณีธรรมคุณ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ก
บทคัดย่อภาษาอังกฤษ ..... ข
กิตติกรรมประกาศ ..... ค
สารบัญ ..... ง
สารบัญตาราง ..... ฉ
สารบัญรูป ..... ช
บทที่ 1 บทนำ 1
1.1 ความสำคัญของปัญหา 1
1.2 วัตถุประสงค์ที่ศึกษา 3
1.3 ประโยชน์ที่คาดว่าจะได้รับ 3
1.4 ขอบเขตการศึกษา 3
1.5 นิยามคำศัพท์เฉพาะ 3
1.6 ประชากรและตัวอย่าง 4
บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง 5
2.1 การทำเหมืองข้อมูล 5
2.2 วิธีนาอีฟเบส์ 6
2.3 วิธีเพื่อนบ้านใกล้สุด k ตัว 7
2.4 วิธีต้นไม้ตัดสินใจ 8
2.5 วิธีโครงข่ายประสาทเทียม 10
2.6 วิธีซัพพอร์ตเวกเตอร์แมชชีน 15
2.7 การเปรียบเทียบประสิทธิภาพของวิธีการจำแนกกลุ่ม 19
2.8 งานวิจัยที่เกี่ยวข้อง 20
บทที่ 3 วิธีการดำเนินงานวิจัย 23
3.1 การเก็บรวบรวมข้อมูล 23
3.2 ตัวแปรและนิยาม 27
3.3 เครื่องมือที่ใช้ในการวิจัย 25
3.4 ขั้นตอนในการดำเนินงานวิจัย 25
บทที่ 4 ผลการวิเคราะห์ข้อมูล 29
4.1 การวิเคราะห์ข้อมูลเบื้องต้นของข้อมูล 30
4.2 ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ 33
4.3 ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK 45
4.4 ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 57

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ(ต่อ)

หน้า

4.5 ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron 82
4.6 ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO 106
4.7 ผลการเปรียบเทียบประสิทธิภาพของวิธีการจำแนกกลุ่มโดยใช้โปรแกรม SPSS และ WEKA 118
บทที่ 5 สรุปผล อภิปรายผล และข้อเสนอแนะ 127
5.1 สรุปผลการวิจัย 127
5.2 อภิปรายผล 128
5.3 ข้อเสนอแนะ 128
บรรณานุกรม 129
ภาคผนวก ก 132
ภาคผนวก ข 154

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญตาราง

ตารางที่		หน้า
4-1	การเปรียบเทียบสัดส่วนค่าความคลาดเคลื่อนของข้อมูลโรคมะเร็งเต้านมของรัฐวิสคอนซิน	30
4-2	การเปรียบเทียบสัดส่วนค่าความคลาดเคลื่อนของข้อมูลโรคเบาหวานของชาวพิม่า ประเทศอินเดีย	31
4-3	การเปรียบเทียบสัดส่วนค่าความคลาดเคลื่อนของข้อมูลการชำระเงินด้วยบัตรเครดิตของลูกค้า	32
4-4	ตารางชุดข้อมูลโรคมะเร็งเต้านมของรัฐวิสคอนซิน ซึ่งเป็นชุดข้อมูลที่มีค่านอกเกณฑ์ระดับน้อย โดยการสุ่มตัวอย่าง seed 10	118
4-5	ตารางชุดข้อมูลโรคมะเร็งเต้านมของรัฐวิสคอนซิน ซึ่งเป็นชุดข้อมูลที่มีค่านอกเกณฑ์ระดับน้อย โดยการสุ่มตัวอย่าง seed 20	119
4-6	ตารางชุดข้อมูลโรคมะเร็งเต้านมของรัฐวิสคอนซิน ซึ่งเป็นชุดข้อมูลที่มีค่านอกเกณฑ์ระดับน้อย โดยการสุ่มตัวอย่าง seed 30	120
4-7	ตารางชุดข้อมูลโรคเบาหวานของชาวพิม่า ประเทศอินเดีย ซึ่งเป็นชุดข้อมูลที่มีค่านอกเกณฑ์ระดับปานกลาง โดยการสุ่มตัวอย่าง seed 10	121
4-8	ตารางชุดข้อมูลโรคเบาหวานของชาวพิม่า ประเทศอินเดีย ซึ่งเป็นชุดข้อมูลที่มีค่านอกเกณฑ์ระดับปานกลาง โดยการสุ่มตัวอย่าง seed 20	122
4-9	ตารางชุดข้อมูลโรคเบาหวานของชาวพิม่า ประเทศอินเดีย ซึ่งเป็นชุดข้อมูลที่มีค่านอกเกณฑ์ระดับปานกลาง โดยการสุ่มตัวอย่าง seed 30	123
4-10	ตารางชุดข้อมูลการชำระเงินด้วยบัตรเครดิตของลูกค้า ซึ่งเป็นข้อมูลที่มีค่านอกเกณฑ์ระดับมาก โดยการสุ่มตัวอย่าง seed 10	124
4-11	ตารางชุดข้อมูลการชำระเงินด้วยบัตรเครดิตของลูกค้า ซึ่งเป็นข้อมูลที่มีค่านอกเกณฑ์ระดับมาก โดยการสุ่มตัวอย่าง seed 20	125
4-12	ตารางชุดข้อมูลการชำระเงินด้วยบัตรเครดิตของลูกค้า ซึ่งเป็นข้อมูลที่มีค่านอกเกณฑ์ระดับมาก โดยการสุ่มตัวอย่าง seed 30	126

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญรูป

รูปที่		หน้า
2-1	เหตุการณ์ E บนเหตุการณ์ k เหตุการณ์ที่เกิดขึ้นพร้อมกันไม่ได้	7
2-2	ตัวอย่างวิธีเพื่อนใกล้สุด k ตัว	8
2-3	ส่วนประกอบของต้นไม้ตัดสินใจ	9
2-4	โครงข่ายของเซลล์ประสาท	11
2-5	ลักษณะโครงข่ายประสาทเทียมแบบส่งสัญญาณไปข้างหน้า	12
2-6	ลักษณะโครงข่ายประสาทเทียมแบบย้อนกลับ	13
2-7	ลักษณะโครงข่ายประสาทเทียมแบบเพอร์เซปตรอนหลายชั้น	13
2-8	การขยายตัวของเส้นขอบ	15
2-9	เส้นขอบและเส้นแบ่งเมื่อแทนด้วยสมการเส้นตรง	16
2-10	รูปแบบการวางตัวที่ไม่สามารถแบ่งด้วยเส้นตรงได้	18
2-11	เมตริกซ์ความสับสน	19
4-1	ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 10 โดยโปรแกรม SPSS	33
4-2	ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 20 โดยโปรแกรม SPSS	33
4-3	ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 30 โดยโปรแกรม SPSS	34
4-4	ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 10 โดยโปรแกรม SPSS	35
4-5	ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 20 โดยโปรแกรม SPSS	35
4-6	ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 30 โดยโปรแกรม SPSS	36
4-7	ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 10 โดยโปรแกรม SPSS	37
4-8	ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 20 โดยโปรแกรม SPSS	37
4-9	ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 30 โดยโปรแกรม SPSS	38
4-10	ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 10 โดยโปรแกรม WEKA	39
4-11	ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 20 โดยโปรแกรม WEKA	39

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญรูป(ต่อ)

รูปที่		หน้า
4-12	ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 30 โดยโปรแกรม WEKA	40
4-13	ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 10 โดยโปรแกรม WEKA	41
4-14	ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 20 โดยโปรแกรม WEKA	41
4-15	ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 30 โดยโปรแกรม WEKA	42
4-16	ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 10 โดยโปรแกรม WEKA	43
4-17	ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 20 โดยโปรแกรม WEKA	43
4-18	ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 30 โดยโปรแกรม WEKA	44
4-19	ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 10 โดยโปรแกรม SPSS	45
4-20	ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 20 โดยโปรแกรม SPSS	45
4-21	ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 30 โดยโปรแกรม SPSS	46
4-22	ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 10 โดยโปรแกรม SPSS	47
4-23	ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 20 โดยโปรแกรม SPSS	47
4-24	ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 30 โดยโปรแกรม SPSS	48

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญรูป(ต่อ)

รูปที่		หน้า
4-25	ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 10 โดยโปรแกรม SPSS	49
4-26	ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 20 โดยโปรแกรม SPSS	49
4-27	ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 30 โดยโปรแกรม SPSS	50
4-28	ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 10 โดยโปรแกรม WEKA	51
4-29	ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 20 โดยโปรแกรม WEKA	51
4-30	ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 30 โดยโปรแกรม WEKA	52
4-31	ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 10 โดยโปรแกรม WEKA	53
4-32	ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึมIBK โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 20 โดยโปรแกรม WEKA	53
4-33	ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 30 โดยโปรแกรม WEKA	54
4-34	ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 10 โดยโปรแกรม WEKA	55
4-35	ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 20 โดยโปรแกรม WEKA	55

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญรูป(ต่อ)

รูปที่		หน้า
4-36	ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 30 โดยโปรแกรม WEKA	56
4-37	ต้นไม้เพื่อการตัดสินใจ โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 10 โดยโปรแกรม SPSS	57
4-38	ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 10 โดยโปรแกรม SPSS	58
4-39	ต้นไม้เพื่อการตัดสินใจ โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 20 โดยโปรแกรม SPSS	58
4-40	ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 20 โดยโปรแกรม SPSS	59
4-41	ต้นไม้เพื่อการตัดสินใจ โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 30 โดยโปรแกรม SPSS	59
4-42	ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 30 โดยโปรแกรม SPSS	60
4-43	ต้นไม้เพื่อการตัดสินใจ โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 10 โดยโปรแกรม SPSS	61
4-44	ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 10 โดยโปรแกรม SPSS	62
4-45	ต้นไม้เพื่อการตัดสินใจ โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 20 โดยโปรแกรม SPSS	62
4-46	ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 20 โดยโปรแกรม SPSS	63
4-47	ต้นไม้เพื่อการตัดสินใจ โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 30 โดยโปรแกรม SPSS	64
4-48	ต้นไม้เพื่อการตัดสินใจ โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 30 โดยโปรแกรม SPSS	65

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญรูป(ต่อ)

รูปที่		หน้า
4-49	ต้นไม้เพื่อการตัดสินใจ การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 10 โดยโปรแกรม SPSS	66
4-50	ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 10 โดยโปรแกรม SPSS	67
4-51	ต้นไม้เพื่อการตัดสินใจ การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 20 โดยโปรแกรม SPSS	67
4-52	ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 20 โดยโปรแกรม SPSS	68
4-53	ต้นไม้เพื่อการตัดสินใจ การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 30 โดยโปรแกรม SPSS	69
4-54	ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 30 โดยโปรแกรม SPSS	69
4-55	ต้นไม้เพื่อการตัดสินใจ โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 10 โดยโปรแกรม WEKA	70
4-56	ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 10 โดยโปรแกรม WEKA	71
4-57	ต้นไม้เพื่อการตัดสินใจ โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 20 โดยโปรแกรม WEKA	71
4-58	ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 20 โดยโปรแกรม WEKA	72
4-59	ต้นไม้เพื่อการตัดสินใจ โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 30 โดยโปรแกรม WEKA	72
4-60	ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 30 โดยโปรแกรม WEKA	73
4-61	ต้นไม้เพื่อการตัดสินใจ โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 10 โดยโปรแกรม WEKA	73

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญรูป(ต่อ)

รูปที่		หน้า
4-62	ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 10 โดยโปรแกรม WEKA	74
4-63	ต้นไม้เพื่อการตัดสินใจ โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 20 โดยโปรแกรม WEKA	74
4-64	ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 20 โดยโปรแกรม WEKA	75
4-65	ต้นไม้เพื่อการตัดสินใจ โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 30 โดยโปรแกรม WEKA	76
4-66	ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 30 โดยโปรแกรม WEKA	76
4-67	ต้นไม้เพื่อการตัดสินใจ การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 10 โดยโปรแกรม WEKA	77
4-68	ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 10 โดยโปรแกรม WEKA	78
4-69	ต้นไม้เพื่อการตัดสินใจ การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 20 โดยโปรแกรม WEKA	78
4-70	ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 20 โดยโปรแกรม WEKA	79
4-71	ต้นไม้เพื่อการตัดสินใจ การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 30 โดยโปรแกรม WEKA	80
4-72	ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 30 โดยโปรแกรม WEKA	81
4-73	โครงข่ายประสาท โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 10 โดยโปรแกรม SPSS	82
4-74	ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 10 โดยโปรแกรม SPSS	83

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญรูป(ต่อ)

รูปที่		หน้า
4-75	โครงข่ายประสาท โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 20 โดยโปรแกรม SPSS	83
4-76	ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 20 โดยโปรแกรม SPSS	84
4-77	โครงข่ายประสาท โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 30 โดยโปรแกรม SPSS	85
4-78	ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 30 โดยโปรแกรม SPSS	85
4-79	โครงข่ายประสาท โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 10 โดยโปรแกรม SPSS	86
4-80	ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 10 โดยโปรแกรม SPSS	87
4-81	โครงข่ายประสาท โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 20 โดยโปรแกรม SPSS	87
4-82	ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 20 โดยโปรแกรม SPSS	88
4-83	โครงข่ายประสาท โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 30 โดยโปรแกรม SPSS	88
4-84	ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 30 โดยโปรแกรม SPSS	89
4-85	โครงข่ายประสาท การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 10 โดยโปรแกรม SPSS	90
4-86	ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 10 โดยโปรแกรม SPSS	91
4-87	โครงข่ายประสาท การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 20 โดยโปรแกรม SPSS	91

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญรูป(ต่อ)

รูปที่		หน้า
4-88	ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 20 โดยโปรแกรม SPSS	92
4-89	โครงข่ายประสาท การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 30 โดยโปรแกรม SPSS	92
4-90	ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 30 โดยโปรแกรม SPSS	93
4-91	โครงข่ายประสาท โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 10 โดยโปรแกรม WEKA	94
4-92	ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 10 โดยโปรแกรม WEKA	95
4-93	โครงข่ายประสาท โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 20 โดยโปรแกรม WEKA	95
4-94	ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 20 โดยโปรแกรม WEKA	96
4-95	โครงข่ายประสาท โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 30 โดยโปรแกรม WEKA	96
4-96	ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 30 โดยโปรแกรม WEKA	97
4-97	โครงข่ายประสาท โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 10 โดยโปรแกรม WEKA	98
4-98	ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 10 โดยโปรแกรม WEKA	99
4-99	โครงข่ายประสาท โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 20 โดยโปรแกรม WEKA	99
4-100	ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 20 โดยโปรแกรม WEKA	100

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญรูป(ต่อ)

รูปที่		หน้า
4-101	โครงข่ายประสาท โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 30 โดยโปรแกรม WEKA	100
4-102	ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 20 โดยโปรแกรม WEKA	101
4-103	โครงข่ายประสาท การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 10 โดยโปรแกรม WEKA	102
4-104	ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 10 โดยโปรแกรม WEKA	103
4-105	โครงข่ายประสาท การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 20 โดยโปรแกรม WEKA	103
4-106	ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 20 โดยโปรแกรม WEKA	104
4-107	โครงข่ายประสาท การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 30 โดยโปรแกรม WEKA	104
4-108	ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 30 โดยโปรแกรม WEKA	105
4-109	ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 10 โดยโปรแกรม SPSS	106
4-110	ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 20 โดยโปรแกรม SPSS	106
4-111	ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 30 โดยโปรแกรม SPSS	107
4-112	ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 10 โดยโปรแกรม SPSS	108

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญรูป(ต่อ)

รูปที่		หน้า
4-113	ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 20 โดยโปรแกรม SPSS	108
4-114	ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 30 โดยโปรแกรม SPSS	109
4-115	ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 10 โดยโปรแกรม SPSS	110
4-116	ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 20 โดยโปรแกรม SPSS	110
4-117	ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 30 โดยโปรแกรม SPSS	111
4-118	ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 10 โดยโปรแกรม WEKA	112
4-119	ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 20 โดยโปรแกรม WEKA	112
4-120	ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 30 โดยโปรแกรม WEKA	113
4-121	ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 10 โดยโปรแกรม WEKA	114
4-122	ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 20 โดยโปรแกรม WEKA	114

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญรูป(ต่อ)

รูปที่		หน้า
4-123	ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 30 โดยโปรแกรม WEKA	115
4-124	ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 10 โดยโปรแกรม WEKA	116
4-125	ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 20 โดยโปรแกรม WEKA	116
4-126	ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 30 โดยโปรแกรม WEKA	117

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 1
บทนำ

1.1 ความสำคัญของปัญหา

ในปัจจุบันข้อมูลที่มีคุณภาพน่าเชื่อถือมีความสำคัญเป็นอย่างมากในการวิเคราะห์ข้อมูล เพื่อการนำข้อมูลไปใช้ประโยชน์ได้สูงสุด บางครั้งจากข้อมูลจะพบว่าข้อมูลมีค่าที่มากเกินไปหรือน้อยเกินไปแฝงอยู่ ซึ่งเรียกว่าค่านอกเกณฑ์ (Outlier) เป็นค่าที่อยู่ปลายสุดซึ่งตกอยู่ใกล้กับขีดจำกัดของพิสัยข้อมูล การหาค่านอกเกณฑ์มีความสำคัญ เนื่องจากแสดงค่าความคลาดเคลื่อนในข้อมูล ถ้าเรานำข้อมูลที่มีค่านอกเกณฑ์ไปวิเคราะห์จะส่งผลให้เกิดความคลาดเคลื่อนของผลลัพธ์ข้อมูลที่ได้ การกระจายของข้อมูลและค่าเฉลี่ยของข้อมูลไม่ดี ส่งผลให้ไม่เป็นไปตามข้อกำหนดเบื้องต้น (Assumption) และทำให้ไม่สามารถนำข้อมูลไปใช้ประโยชน์ได้อย่างสูงสุด สำหรับสาเหตุที่ทำให้เกิดค่านอกเกณฑ์ ความคลาดเคลื่อนจากการแปรผันข้อมูลที่เก็บรวบรวมมา ซึ่งเป็นความคลาดเคลื่อนที่ไม่สามารถควบคุมได้ ความคลาดเคลื่อนที่เกิดจากเครื่องมือที่ใช้วัดมีคุณภาพต่ำทำให้เกิดค่านอกเกณฑ์ ความคลาดเคลื่อนที่เกิดจากการบันทึกข้อมูลจากการปฏิบัติโดยไม่ตรวจสอบให้ถี่ถ้วน เพื่อให้ได้ผลการวิเคราะห์ที่เชื่อถือได้ การตรวจสอบหาค่านอกเกณฑ์จึงเป็นสิ่งที่สำคัญก่อนการนำข้อมูลไปวิเคราะห์ เพื่อให้ได้ข้อมูลที่มีความผิดปกติและสามารถแก้ไขได้ บางครั้งข้อมูลอาจจะไม่สามารถใช้งานได้เต็มประสิทธิภาพหรือตรงตามความต้องการ เราจึงมีการจำแนกกลุ่มของข้อมูล เพื่อให้ได้วิธีที่มีความเหมาะสมกับข้อมูลที่มีความแตกต่างกันไป ดังนั้นผู้ทำการวิจัยจึงต้องเลือกวิธีการจำแนกกลุ่มให้เหมาะสมกับข้อมูล (วรพรรณ เจริญขำ, 2556)

จากการศึกษางานวิจัยที่เกี่ยวข้อง นิเวศ จิระวิชิตชัย (2553) ได้ทำการศึกษางานวิจัยเกี่ยวกับโรคมะเร็งเต้านม เรื่องการค้นหาเทคนิคเหมืองข้อมูลเพื่อสร้างโมเดลการวิเคราะห์โรคอัตโนมัติ งานวิจัยนี้มุ่งเน้นการค้นหาเทคนิคด้านเหมืองข้อมูล สร้างโมเดลการวิเคราะห์โรคอัตโนมัติ ค้นหาอัลกอริทึมที่เหมาะสมที่สุดสำหรับฐานข้อมูลทางการแพทย์ โดยวัดประสิทธิภาพจากค่าความถูกต้อง พบว่าวิธีต้นไม้ตัดสินใจให้ประสิทธิภาพค่าความถูกต้องดีที่สุดคือ 75.52% เมื่อเปรียบเทียบกับวิธีซัพพอร์ตเวกเตอร์แมชชีนและวิธีเพื่อนบ้านใกล้สุด k ตัว Sriwiboon, N. (2016) ศึกษางานวิจัยเกี่ยวกับโรคมะเร็ง เรื่องการเปรียบเทียบประสิทธิภาพอัลกอริทึมเหมืองข้อมูลเพื่อวิเคราะห์ปัจจัยที่ส่งผลต่อการเกิดโรคมะเร็ง งานวิจัยนี้นำเทคนิคการทำเหมืองข้อมูลมาประยุกต์ใช้กับการตรวจวิเคราะห์การเกิดโรคมะเร็ง เพื่อนำกฎการจำแนกข้อมูลที่ได้ไปพัฒนาเป็นระบบตรวจวิเคราะห์ปัจจัยที่ส่งผลต่อการเกิดโรคมะเร็ง โดยวัดประสิทธิภาพจากค่าสัมบูรณ์ของความคลาดเคลื่อนเฉลี่ย พบว่าวิธีต้นไม้ตัดสินใจให้ค่าความถูกต้องสูงสุดคือ 98.63% เมื่อเปรียบเทียบกับวิธีเพื่อนบ้านใกล้สุด k ตัวและวิธีนาอีฟเบส์ Priya, R. and Aruna, P. (2012) ศึกษางานวิจัยเกี่ยวกับโรคเบาหวาน เรื่องวิธีซัพพอร์ตเวกเตอร์แมชชีนและวิธีโครงข่ายประสาทเทียมในการวินิจฉัยโรคเบาหวานในจอประสาทตา ซึ่งเป็นโรคที่มีสาเหตุเกิดจากโรคแทรกซ้อนของโรคเบาหวานโดยจำแนกกลุ่มออกเป็น 2 กลุ่ม คือผู้ที่เป็นโรคเบาหวานในจอประสาทตาและไม่เป็นโรคเบาหวานในจอประสาทตา พบว่าวิธีซัพพอร์ตเวกเตอร์แมชชีนให้ค่า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ความถูกต้อง 97.61% มากกว่าวิธีโครงข่ายประสาทเทียม กิตติพล วิแสงและคณะ (2552) ศึกษางานวิจัยเกี่ยวกับโรคเบาหวาน เรื่องการวิเคราะห์ปัจจัยเสี่ยงของโรคเบาหวาน เพื่อนำมาเป็นเครื่องมือประเมินการเกิดโรคเบาหวานโดยไม่ต้องอาศัยการตรวจเลือด พบว่าวิธีโครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับให้ค่าความถูกต้องมากที่สุด เมื่อเปรียบเทียบกับวิธีโครงข่ายประสาทเทียมแบบธรรมดาและวิธีนาอีฟเบส์ เดช ธรรมศิริ (2552) ศึกษางานวิจัยเกี่ยวกับการให้คะแนนสินเชื่อ เรื่องการให้คะแนนสินเชื่อโดยวิธีการทำเหมืองข้อมูลด้วยวิธีซับพอร์ตเวกเตอร์แมชชีนรวมทั้งการเลือกใช้ลักษณะที่เหมาะสมร่วมกับการหาค่าพารามิเตอร์ที่เหมาะสมด้วยวิธีค้นหาแบบกริช เพื่อลดความเสี่ยงสำหรับการให้เครดิตสินเชื่อแก่ลูกค้าที่มีความเสี่ยงในการผิดสัญญาหรือขาดการชำระเงินในการหาโมเดลที่เหมาะสม พบว่าวิธีต้นไม้ตัดสินใจมีค่าความถูกต้องมากที่สุดคือ 79.48% เมื่อเปรียบเทียบกับวิธีโครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับและวิธีซัพพอร์ตเวกเตอร์แมชชีน ทิพย์ธิดา วงศ์พิพันธ์ (2555) ศึกษางานวิจัยเกี่ยวกับสินเชื่อ เรื่องการใช้เหมืองข้อมูลช่วยในการตัดสินใจการให้สินเชื่อ เพื่อเป็นแนวทางการสนับสนุนการตัดสินใจการอนุมัติสินเชื่อของบริษัทได้อย่างมีประสิทธิภาพมากขึ้น ซึ่งมีส่วนช่วยลดปริมาณหนี้สินสูญได้ พบว่าวิธีต้นไม้ตัดสินใจให้ค่าความถูกต้องมากที่สุดคือ 90.47% มากกว่าวิธีนาอีฟเบส์

ดังนั้นปัญหาพิเศษฉบับนี้ทำการศึกษาการจำแนกกลุ่ม (Classification) ด้วยวิธีต่างๆ 5 วิธี คือ วิธีนาอีฟเบส์ (Naive Bayes) วิธีเพื่อนบ้านใกล้สุด k ตัว (k-Nearest Neighbor) วิธีต้นไม้ตัดสินใจ (Decision Tree) วิธีโครงข่ายประสาทเทียม (Artificial Neural Network) และวิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) เพื่อต้องการเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มทั้ง 5 วิธี ว่าวิธีใดมีประสิทธิภาพและเหมาะสมกับรูปแบบของชุดข้อมูล โดยใช้วิธีการสุ่มตัวอย่างระหว่างโปรแกรม SPSS กับ WEKA เพื่อเปรียบเทียบว่าโปรแกรมใดให้ค่าความถูกต้องสูงกว่า ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยและค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ยต่ำกว่า

1.2 วัตถุประสงค์ที่ศึกษา

1. เพื่อเปรียบเทียบประสิทธิภาพในการจำแนกกลุ่ม 5 วิธี คือ วิธีนาอีฟเบส์ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม และวิธีซัพพอร์ตเวกเตอร์แมชชีน ว่าวิธีใดให้ค่าความถูกต้อง (Accuracy) ของการพยากรณ์สูงกว่า ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error : MSE) ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย (Mean Absolute Deviation : MAD) ต่ำกว่า

2. เพื่อเปรียบเทียบวิธีการสุ่มตัวอย่างระหว่างโปรแกรม SPSS กับ WEKA ว่าโปรแกรมใดให้ค่าความถูกต้องสูงกว่า ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยและค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ยต่ำกว่า

1.3 ประโยชน์ที่คาดว่าจะได้รับ

สามารถนำผลการเปรียบเทียบประสิทธิภาพที่ได้ไปใช้เป็นแนวทางในการเลือกวิธีการจำแนกกลุ่มที่เหมาะสมที่สุดและช่วยลดข้อผิดพลาดในการเลือกใช้วิธีการจำแนกกลุ่ม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

1.4 ขอบเขตการศึกษา

ในการทำปัญหาพิเศษได้ทำการแบ่งชุดข้อมูลอย่างสุ่มด้วยโปรแกรม SPSS กับ WEKA ออกเป็น 3 ส่วน โดยมีอัตราส่วน 70:20:10 (พยูน พาณิชย์กุล, 2548) ในข้อมูลส่วนที่ 1 ข้อมูลเรียนรู้ (Training data) นำไปสร้างตัวแบบ (Model) 70 เปอร์เซ็นต์ ข้อมูลส่วนที่ 2 ข้อมูลตรวจสอบความถูกต้อง (Validation data) นำไปประเมินความผิดพลาดของตัวแบบ 20 เปอร์เซ็นต์ และข้อมูลส่วนที่ 3 ข้อมูลทดสอบ (Testing data) นำไปทดสอบตัวแบบ 10 เปอร์เซ็นต์ โดยการกำหนดตัวเลขสุ่มเทียม (Random Seed) เป็น 10, 20, 30 ซึ่งจะใช้วิธีในการทดสอบประกอบด้วย 5 วิธี คือ วิธีนาอีฟเบส์ วิธีเพื่อนบ้านใกล้สุด k ตัว ใช้อัลกอริทึมชนิด IBK วิธีต้นไม้ตัดสินใจ ใช้อัลกอริทึมชนิด J48 (C4.5) วิธีโครงข่ายประสาทเทียม ใช้อัลกอริทึมชนิดเพอร์เซปตรอนแบบหลายชั้น โดยกำหนดค่าอัตราการเรียนรู้เป็น 0.1 ค่าโมเมนตัมเป็น 0.9 (A. Berson and J. Smith, 2001) จำนวนรอบการสอน 20,000 รอบ และชั้นซ่อน 1 ชั้น (สายชล สินสมบูรณ์ทอง, 2560) และวิธีซัพพอร์ตเวกเตอร์แมชชีน ใช้วิธีอัลกอริทึม SMO ชนิดโพลิโนเมียลเคอร์เนล ในการเปรียบเทียบประสิทธิภาพของการจำแนกกลุ่มทั้ง 5 วิธี ใช้ค่าความถูกต้อง ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย และค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย ทำการเก็บรวบรวมข้อมูลเป็นจำนวน 3 ชุด คือ โรคมะเร็งเต้านมของรัฐวิสคอนซิน (Breast Cancer Wisconsin) โรคเบาหวานของชาวพิม่า ประเทศอินเดีย (Pima Indians Diabetes) และการชำระเงินด้วยบัตรเครดิตของลูกค้า (Default of Credit Card Clients)

1.5 นิยามคำศัพท์เฉพาะ

ค่านอกเกณฑ์ (Outlier) หมายถึง ค่าที่อยู่ปลายสุดซึ่งตกอยู่ใกล้กับขีดจำกัดของพิสัยข้อมูล การหาค่าผิดปกติมีความสำคัญเนื่องจากค่านอกเกณฑ์อาจจะแสดงความคลาดเคลื่อนในข้อมูล แม้ว่าค่านอกเกณฑ์เป็นข้อมูลที่ถูกต้องและไม่มีความคลาดเคลื่อน (สายชล สินสมบูรณ์ทอง, 2558) ค่านอกเกณฑ์คือ ค่าสังเกตที่เบี่ยงเบนไปจากค่าสังเกตอื่นมาก จนทำให้สงสัยว่าค่าสังเกตนั้นได้มาจากวิธีอื่น ลักษณะของค่านอกเกณฑ์มี 2 ลักษณะ (Hawkins, 1980) คือ ค่าสังเกตที่มีค่าสูงหรือต่ำมาก หรือเป็นค่าที่เบี่ยงเบนไปจากค่าสังเกตส่วนใหญ่เรียกค่านอกเกณฑ์นี้ว่า Discordant observations และค่าสังเกตที่มีลักษณะการแจกแจงแตกต่างจากลักษณะการแจกแจงของประชากรที่สนใจศึกษา เรียกค่านอกเกณฑ์นี้ว่า Contaminate observations
(R. J. Beckman and R . D. Cook, 1983)

มะเร็งเต้านม (Breast Cancer) เกิดจากความผิดปกติของเซลล์ที่อยู่ภายในท่อน้ำนมหรือต่อมน้ำนม เซลล์เหล่านี้มีการแบ่งตัวผิดปกติ ไม่สามารถควบคุมได้ มักแพร่กระจายไปตามทางเดินน้ำเหลืองไปสู่อวัยวะที่ใกล้เคียง เช่น ต่อมน้ำเหลืองที่รักแร้ หรือแพร่กระจายไปสู่อวัยวะที่อยู่ห่างไกล เช่น กระดูก ปอด ตับ และสมอง เช่นเดียวกับมะเร็งชนิดอื่นๆ เมื่อเซลล์มะเร็งมีจำนวนมากขึ้นก็จะแย่งสารอาหารและปล่อยสารบางอย่างที่เป็นอันตรายและทำลายอวัยวะต่างๆ จนทำให้ผู้ป่วยเสียชีวิตในที่สุด (คณะแพทยศาสตร์ศิริราชพยาบาล, 2555)

เบาหวาน (Diabetes Mellitus หรือ Diabetes หรือเรียกย่อๆว่า โรคดีเอ็ม) เป็นโรคเรื้อรังที่รักษาไม่หาย ต้องได้รับการดูแลรักษาตลอดชีวิต ทั้งนี้เกิดจากการที่ในเลือดมีน้ำตาลสูงกว่าปกติ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งไม่อนุญาตให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

โดยทั่วไปแบ่งโรคเบาหวานได้เป็น 3 ชนิด คือ โรคเบาหวานชนิดต้องพึ่งอินซูลิน โรคเบาหวานชนิดที่ไม่จำเป็นต้องพึ่งอินซูลิน และโรคเบาหวานในหญิงตั้งครรภ์ (พวงทอง ไกรพิบูลย์, 2557)

บัตรเครดิต (Credit Card) เป็นบัตรพลาสติกที่บนบัตรมีข้อมูลของผู้ใช้บัตร เช่น ลายเซ็น ชื่อ เป็นต้น บัตรเครดิตจะให้อำนาจแก่ผู้ถือบัตรในการซื้อสินค้าหรือบริการต่างๆ โดยค่าใช้จ่ายในการซื้อสินค้าหรือบริการนั้นๆ จะถูกเรียกเก็บในภายหลัง ในปัจจุบันบัตรเครดิตสามารถใช้ได้ผ่านตู้ ATM ร้านค้าต่างๆ ธนาคาร และการชำระเงินออนไลน์ (2558 : ออนไลน์)

ตัวสร้างเลขสุ่มเทียม (Random Seed) หรือตัวสร้างบิตสุ่มแบบกำหนดได้ (Deterministic random bit generator: DRBG) เป็นขั้นตอนวิธีสำหรับใช้ในการสร้างลำดับของตัวเลขที่มีความใกล้เคียงกับสมบัติของการสุ่ม ถึงแม้ว่าลำดับตัวเลขที่ได้จากขั้นตอนวิธีตัวสร้างเลขสุ่มเทียมนี้จะใกล้เคียงกับลำดับเลขสุ่มแท้จริงมากแค่ไหน แต่ก็ไม่ได้เป็นลำดับตัวเลขแบบสุ่มที่แท้จริงเนื่องจากลำดับตัวเลขที่ได้ออกมาจากตัวสร้างเลขสุ่มเทียมทั้งหมดได้มาจากกลุ่มเล็กๆ ของค่าเริ่มต้นที่เรากำหนดให้เป็นตัวตั้งต้น (seed) ของตัวสร้างเลขสุ่มเทียม (2560 : ออนไลน์)

1.6 ประชากรและตัวอย่าง

1.6.1 โรคมะเร็งเต้านมของรัฐวิสคอนซิน (Breast Cancer Wisconsin)

ประชากร คือ ผู้ป่วยที่เป็นมะเร็งเต้านมของรัฐวิสคอนซิน

ตัวอย่าง คือ ผู้ป่วยที่เป็นโรคมะเร็งเต้านมจากโรงพยาบาลเมดิสัน รัฐวิสคอนซิน ล่าสุดจนถึงวันที่ 8 มกราคม ค.ศ.1991

1.6.2 โรคเบาหวานของชาวพิม่า ประเทศอินเดีย (Pima Indians Diabetes)

ประชากร คือ ผู้ป่วยที่เป็นโรคเบาหวาน ชาวพิม่า ประเทศอินเดีย

ตัวอย่าง คือ ผู้ป่วยที่เป็นโรคเบาหวาน ชาวพิม่า ประเทศอินเดีย จากสถาบันแห่งชาติของโรคเบาหวาน ทางเดินอาหารและโรคไต ล่าสุดจนถึงวันที่ 9 พฤษภาคม ค.ศ.1990

1.6.3 การชำระเงินด้วยบัตรเครดิตของลูกค้า (Default of Credit Card Clients)

ประชากร คือ ลูกค้าบัตรเครดิตของประเทศไต้หวัน

ตัวอย่าง คือ การชำระเงินด้วยบัตรเครดิตของลูกค้า จากสถาบันภาควิชาการจัดการเทคโนโลยีสารสนเทศ มหาวิทยาลัยวงฮา ประเทศไต้หวัน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้ ผู้วิจัยได้นำเสนอเนื้อหาที่เน้นถึงทฤษฎีและงานวิจัยที่เกี่ยวข้องโดยมีรายละเอียดของเนื้อหาประกอบด้วยหัวข้อย่อย 8 หัวข้อ ดังนี้

2.1 การทำเหมืองข้อมูล (Data Mining)

2.2 วิธีนาอีฟเบส์ (Naive Bayes)

2.3 วิธีเพื่อนบ้านใกล้สุด k ตัว (k-Nearest Neighbor)

2.4 วิธีต้นไม้ตัดสินใจ (Decision Tree)

2.5 วิธีโครงข่ายประสาทเทียม (Artificial Neural Network)

2.6 วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)

2.7 การเปรียบเทียบประสิทธิภาพของวิธีการจำแนกกลุ่ม

2.8 งานวิจัยที่เกี่ยวข้อง

2.1 การทำเหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูลคือการวิเคราะห์ข้อมูลเพื่อแยกประเภท จำแนกรูปแบบและความสัมพันธ์ของข้อมูลจากฐานข้อมูลที่มีขนาดใหญ่หรือคลังข้อมูล โดยมีวิธีต่างๆ หลายวิธี ซึ่งรูปแบบการทำเหมืองข้อมูลนั้นได้รวบรวมความรู้จากหลายแขนงเข้าไว้ด้วยกันซึ่งประกอบด้วยระบบการเรียนรู้ของเครื่องจักร (Machine Learning) รวมกับวิทยาศาสตร์สารสนเทศ (Information Sciences) สถิติ (Statistic) และระบบฐานข้อมูล (Database System) โดยทั่วไปแล้ววิธีที่นำมาใช้ส่วนใหญ่มี 5 ประเภท (รุจิรา ธรรมสมบัติ, 2554)

2.1.1 วิธีการจำแนกกลุ่ม (Classification)

เป็นวิธีในการจำแนกกลุ่มข้อมูลด้วยคุณลักษณะต่างๆ ที่ได้มีการกำหนดไว้แล้ว วิธีนี้เหมาะกับการสร้างตัวแบบเพื่อการพยากรณ์ค่าข้อมูล (Predictive Modeling) ในอนาคตจากการที่ได้จำแนกกลุ่มข้อมูลตัวอย่างไว้แล้ว ซึ่งในลักษณะดังกล่าวเรียกว่า การเรียนรู้แบบมีผู้สอน (Supervised Learning) วิธีการจำแนกกลุ่มเป็นกระบวนการสร้างตัวแบบเพื่อจัดข้อมูลให้อยู่ในกลุ่มที่กำหนด ตัวอย่างเช่น การแบ่งประเภทลูกค้าว่าเชื่อถือได้หรือไม่ ซึ่งเป็นการสร้างตัวแบบโดยการเรียนรู้จากข้อมูลที่ได้กำหนดไว้เรียบร้อยแล้ว

2.1.2 วิธีการค้นหากฎความสัมพันธ์ (Association Rule Discovery)

เป็นวิธีที่ใช้ในการค้นหาความสัมพันธ์ของฐานข้อมูลที่มีขนาดใหญ่ เพื่อที่จะทำการวิเคราะห์ข้อมูลและหาสิ่งที่ซ่อนอยู่ในข้อมูลนั้น เช่น การวิเคราะห์ข้อมูลการซื้อขายในซุปเปอร์มาร์เก็ต

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เพื่อทำการวางแผนการส่งเสริมการขาย (Promotion) และเตรียมการวางแผนการเรียงชั้นวางสินค้า (Shelf) เช่น การวางน้ำอัดลมกับข้าวโพดคั่วไว้ใกล้กัน

2.1.3 วิธีการจัดกลุ่ม (Clustering)

เป็นวิธีการลดขนาดของข้อมูลด้วยการรวมกลุ่มตัวแปรที่มีลักษณะเดียวกันไว้ด้วยกัน ทำให้สามารถค้นหาข้อมูลที่ถูกละเลยไปได้ วิธีนี้มักถูกใช้เป็นขั้นตอนเบื้องต้นในการทำเหมืองข้อมูล และเหมาะกับข้อมูลที่ยังไม่มีกลุ่มอย่างชัดเจน จึงทำการรวมกลุ่มเพื่อหากลุ่มต่างๆ ของข้อมูล โดยจำนวนกลุ่มของข้อมูลแทนด้วย k ซึ่งผู้ที่ใช้วิธีนี้จะเป็นผู้กำหนดจำนวนกลุ่ม วิธีนี้อาจเรียกว่าการจัดกลุ่มแบบเฉลี่ย k กลุ่ม (k-mean Clustering)

2.1.4 วิธีการหาค่าที่แตกต่างจากค่ามาตรฐาน (Deviation Detection)

เป็นวิธีในการหาค่าที่แตกต่างไปจากค่ามาตรฐาน หรือค่าที่คาดคิดไว้ว่าต่างไปมากน้อยเพียงใด โดยทั่วไปมักใช้วิธีทางสถิติหรือการแสดงให้เห็นภาพ สำหรับวิธีนี้ใช้ในการตรวจสอบลายเซ็นหรือปลอมบัตรเครดิต เป็นต้น

2.1.5 วิธีการวิเคราะห์ลำดับ (Sequential Analysis)

เป็นวิธีในการวิเคราะห์ลำดับเพื่อค้นหารูปแบบของการปรากฏของข้อมูล ซึ่งปรากฏในรายการที่แยกออกมา เช่น ถ้าผู้ซื้อสินค้า A แล้วเขาจะซื้อสินค้า B ในภายหลัง วิธีนี้จะแตกต่างจากวิธีการค้นหาความกฎสัมพันธ์ เพราะคำนึงถึงลำดับการซื้อด้วย

2.2 วิธีนาอีฟเบส์ (Naive Bayes)

นาอีฟเบส์เป็นเครื่องจักรเรียนรู้ที่อาศัยความน่าจะเป็น (Probability) ตามทฤษฎีของเบส์ (Bayes Theorem) ซึ่งมีอัลกอริทึมที่ไม่ซับซ้อน เป็นการจำแนกข้อมูลโดยการเรียนรู้ปัญหาที่เกิดขึ้นเพื่อนำมาสร้างเงื่อนไขการจำแนกข้อมูลใหม่ เป็นการจำแนกข้อมูลโดยใช้ความน่าจะเป็นและคำนวณการแจกแจงความน่าจะเป็นตามสมมติฐานที่ตั้งให้กับข้อมูล จากการคำนวณตัวอย่างใหม่ที่ได้จะถูกนำมาปรับเปลี่ยนการแจกแจงซึ่งมีผลต่อการเพิ่มหรือลดความน่าจะเป็นของข้อมูลใหม่ที่เกิดขึ้นและตัวแบบที่ตั้งไว้ให้กับข้อมูลจะถูกปรับเปลี่ยนไปตามข้อมูลใหม่ โดยผนวกกับข้อมูลเดิมที่มีหลักการของนาอีฟเบส์ ใช้การคำนวณหาความน่าจะเป็นซึ่งถูกใช้ในการทำนายผล เป็นวิธีในการแก้ปัญหาแบบจำแนกที่สามารถคาดการณ์ผลลัพธ์ จะทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์ นาอีฟเบส์เป็นวิธีจำแนกกลุ่มข้อมูลที่มีประสิทธิภาพ มีอัลกอริทึมในการใช้งานที่ไม่ซับซ้อน เหมาะกับกรณีของเซตตัวอย่างที่มีจำนวนมากและคุณลักษณะ (Attribute) ของตัวอย่างไม่ขึ้นต่อกันโดยกำหนดให้ความน่าจะเป็นของข้อมูลเท่ากับสมการ 2-1 (วรรณสิริ ธุระชน และคณะ, 2557)

$$P(a_1, a_2, ..., a_n | v_j) = P(v_j) \prod_{i=1}^{n} P(a_i | v_j) \quad (2\text{-}1)$$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

โดยที่ $P(v_j)$ คือ ค่าความน่าจะเป็นของข้อมูลที่ให้คำตอบ v_j

$P(a_i|v_j)$ คือ ค่าความน่าจะเป็นของข้อมูลคุณลักษณะที่ i มีค่า a_i และให้คำตอบ v_j

$P(a_1,a_2,...,a_n|v_j)$ คือ ค่าความน่าจะเป็นของข้อมูลทดสอบ ที่มีค่าคุณลักษณะ $a_1,a_2,...,a_n$ และจะให้คำตอบ v_j

กลุ่ม v_j สำหรับข้อมูลที่มีสมบัติ n ตัว คือ $x = \{a_1,a_2,...,a_n\}$ ใช้สัญลักษณ์ว่า $P(a_1,a_2,...,a_n|v_j)$ โดยที่ $\prod$ หมายถึง ผลคูณของค่า $(a_i|v_j)$ ทั้งหมด i = 1, 2, 3,..., n และ j = 1, 2 ,3,..., n

กำหนดให้ E แทนข้อมูลที่นำมาใช้ในการคำนวณการแจกแจงความน่าจะเป็นหลัง (Posteriori Probability) ของสมมติฐาน h คือ $P(h|E)$ ตามทฤษฎีของนาอีฟเบส์ ดังสมการ 2-2

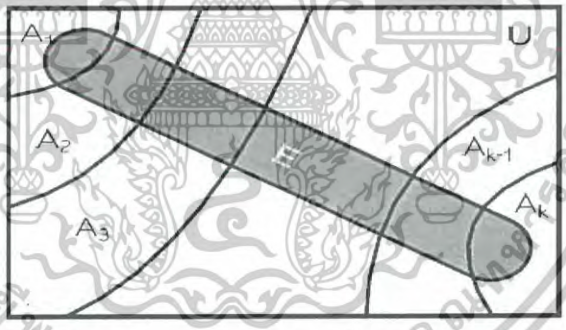
$$P(h|E) = \frac{P(E|h)\,P(h)}{P(E)} \qquad (2\text{-}2)$$

$P(h)$ คือ ความน่าจะเป็นก่อนหน้าของสมมติฐาน h

$P(E)$ คือ ความน่าจะเป็นก่อนหน้าของชุดข้อมูลตัวอย่าง

$P(h|E)$ คือ ความน่าจะเป็นของ h เมื่อรู้ E

$P(E|h)$ คือ ความน่าจะเป็นของ E เมื่อรู้ h



รูปที่ 2-1 เหตุการณ์ E บนเหตุการณ์ k เหตุการณ์ที่เกิดพร้อมกันไม่ได้

2.3 วิธีเพื่อนบ้านใกล้สุด k ตัว (k-Nearest Neighbor)

วิธีเพื่อนบ้านใกล้สุด k ตัวเป็นวิธีการที่ได้รับความนิยมในการใช้งานเป็นอย่างมาก สาเหตุเนื่องจากเป็นวิธีการที่ง่ายและมีประสิทธิภาพ ซึ่งสามารถนำไปประยุกต์ใช้กับงานได้อย่างหลากหลาย เช่น งานทางด้านการจำแนกกลุ่ม (Classification) รวมถึงงานทางด้านการแทนที่ข้อมูลที่สูญหาย (Missing Values Imputation) ซึ่งมีวิธีการดำเนินการดังนี้

(O. Troyanskaya, 2001)

1) กำหนดค่า k เพื่อใช้พิจารณาสมาชิกที่อยู่ใกล้กันมากที่สุด เช่น k=3 คือจะพิจารณาเฉพาะข้อมูล 3 ตัวแรกที่อยู่ใกล้กับจุดที่ต้องการทำนาย

2) คำนวณหาระยะห่างระหว่างข้อมูลตัวอย่างที่สนใจกับข้อมูลอื่นๆ ทุกตัวด้วยระยะห่างยูคลิเดียน (Euclidian distance) จากสมการ 2-3

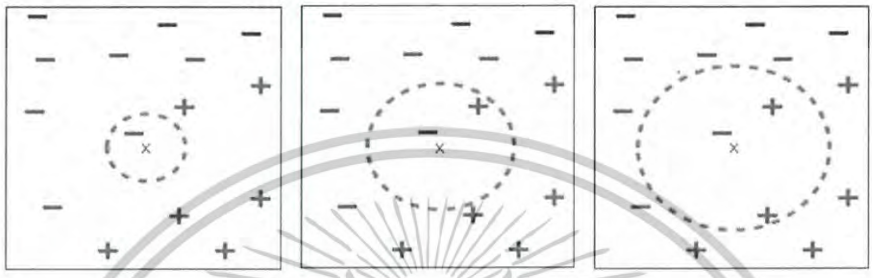
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

$$D_{\text{Euclidian}}(x_i, y_i) = \sqrt{\sum_{k=1}^{n}(x_i, y_i)^2} \quad (2\text{-}3)$$

โดยที่ $D_{\text{Euclidian}}(x_i, y_i)$ คือ ระยะห่างระหว่างตัวอย่าง x_i กับตัวอย่าง y_i

k คือ คุณลักษณะทั้งหมดของตัวอย่าง

3) เลือกค่าข้อมูลที่มีค่าระยะห่างน้อยที่สุด k ตัว เพื่อนำมาพิจารณาหาคำตอบ ดังรูปที่ 2-2



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

(a) วิธีเพื่อนบ้านใกล้สุด k ตัว โดยพิจารณาจากข้อมูล 1 ตัว

(b) วิธีเพื่อนบ้านใกล้สุด k ตัว โดยพิจารณาจากข้อมูล 2 ตัว

(c) วิธีเพื่อนบ้านใกล้สุด k ตัว โดยพิจารณาจากข้อมูล 3 ตัว

รูปที่ 2-2 ตัวอย่างวิธีเพื่อนบ้านใกล้สุด k ตัว

2.4 วิธีต้นไม้ตัดสินใจ (Decision Tree)

เป็นการนำข้อมูลมาสร้างตัวแบบการพยากรณ์ในรูปแบบของโครงสร้างต้นไม้ ซึ่งมีการเรียนรู้ข้อมูลแบบมีผู้สอน (Supervised Learning) สามารถสร้างตัวแบบการจัดกลุ่ม (Clustering) ได้จากกลุ่มตัวอย่างของข้อมูลฝึกหัด (Training Data Set) ได้โดยอัตโนมัติและสามารถพยากรณ์กลุ่มของรายการที่ยังไม่เคยนำมาจัดกลุ่มได้อีกด้วย (รุจิรา ธรรมสมบัติ, 2554)

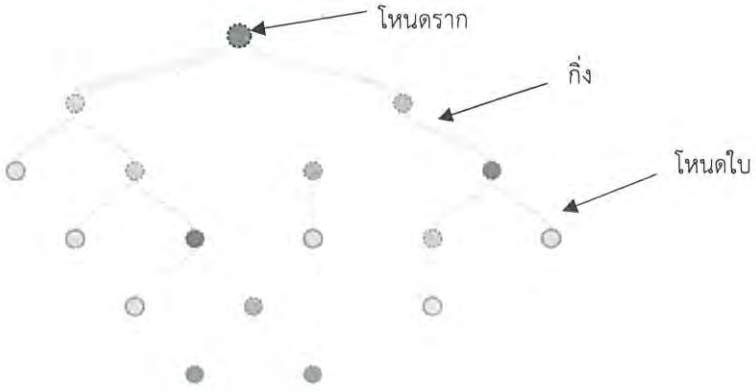
2.4.1 ส่วนประกอบของต้นไม้ตัดสินใจ

1) โหนด (Node) คือ สมบัติต่างๆ เป็นจุดที่แยกข้อมูลว่าจะให้ไปในทิศทางใด ซึ่งโหนดที่อยู่สูงสุดเรียกว่า โหนดราก (Root Node)

2) กิ่ง (Branch) คือคุณสมบัติของโหนดที่แตกออกมา โดยจำนวนของกิ่งจะเท่ากับสมบัติของโหนด

3) ใบ (Leaf) คือ กลุ่มของผลลัพธ์ในการแยกแยะข้อมูล ซึ่งโหนดที่อยู่ล่างสุดเรียกว่าโหนดใบ (Leaf Node) โดยสามารถแสดงส่วนประกอบของต้นไม้ตัดสินใจ ดังรูปที่ 2-3

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 2-3 ส่วนประกอบของต้นไม้ตัดสินใจ

2.4.2 การสร้างต้นไม้ตัดสินใจ

หลักการพื้นฐานของการสร้างต้นไม้ตัดสินใจเป็นการสร้างในลักษณะจากบนลงล่าง (Top–Down) คือเริ่มจากการสร้างรากของต้นไม้ก่อนแล้วจึงแตกกิ่งไปจนถึงใบ โดยแสดงขั้นตอนการสร้างต้นไม้ตัดสินใจได้ดังนี้

1) ต้นไม้เริ่มต้นโดยมีโหนดเพียงโหนดเดียวแสดงถึงชุดข้อมูลฝึกหัด (Training Data Set)

2) ถ้าข้อมูลทั้งหมดอยู่ในกลุ่มเดียวกัน แล้วให้โหนดนั้นเป็นใบและตั้งชื่อแยกตามกลุ่มของข้อมูลนั้น

3) ถ้าในโหนดมีข้อมูลหลายกลุ่มปะปนอยู่ จะต้องวัดค่าผลกำไร (Gain) ของแต่ละคุณลักษณะ (Attribute) เพื่อที่จะเป็นเกณฑ์ในการคัดเลือกคุณลักษณะที่มีความสามารถในการแบ่งแยกข้อมูลออกเป็นกลุ่มต่างๆ ได้ดีที่สุด โดยมีคุณลักษณะที่มีผลกำไรมากที่สุดจะถูกเลือกให้เป็นตัวทดสอบหรือคุณลักษณะที่ใช้ในการตัดสินใจ โดยแสดงในรูปของโหนดบนต้นไม้

4) กิ่งของต้นไม้ถูกสร้างขึ้นจากค่าต่างๆ ที่เป็นไปได้ของโหนดทดสอบ ข้อมูลจะถูกทดสอบและถูกแบ่งออกตามกิ่งต่างๆที่สร้างขึ้น

5) ทำการวนซ้ำเพื่อหาคุณลักษณะที่มีผลกำไรมากที่สุด สำหรับข้อมูลที่ถูกแบ่งแยกออกมาในแต่ละกิ่งเพื่อนำคุณลักษณะนี้มาสร้างเป็นโหนดการตัดสินใจต่อไป โดยที่คุณลักษณะที่ถูกเลือกมาเป็นโหนดแล้ว จะไม่ถูกเลือกมาอีกสำหรับโหนดในระดับต่อไป

6) ทำการวนซ้ำเพื่อแบ่งข้อมูลและแตกกิ่งต้นไม้ไปเรื่อยๆ โดยการวนซ้ำจะสิ้นสุดก็ต่อเมื่อเงื่อนไขข้อใดข้อหนึ่งข้างบนนี้เป็นจริง

2.4.3 การคำนวณค่าข้อมูลผลกำไร (Information Gain)

ต้นไม้ตัดสินใจเป็นโครงสร้างที่ใช้แสดงกฎที่ได้จากเทคนิคการจำแนกประเภทข้อมูลโดยต้นไม้ตัดสินใจจะมีลักษณะคล้ายโครงสร้างต้นไม้ โดยที่แต่ละโหนดแสดงคุณลักษณะ ในการสร้างต้นไม้

เอกสารนี้มีไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตัดสินใจ ปัญหาสำคัญที่ต้องพิจารณาคือควรจะตัดสินใจเลือกคุณลักษณะใดทำหน้าที่เป็นโหนดราก ในแต่ละขั้นตอนในการสร้างต้นไม้และต้นไม้ย่อย (Subtree) ของต้นไม้ตัดสินใจ เกณฑ์ที่ช่วยประกอบการเลือกคุณลักษณะคือค่าเกณฑ์ผลกำไร (Gain Criterion) ซึ่งเป็นค่าที่บ่งบอกคุณลักษณะนั้นสามารถจำแนกกลุ่มของข้อมูลได้ดีเพียงใด โดยทดลองเลือกแต่ละคุณลักษณะที่เป็นไปได้จากชุดข้อมูลมาทำหน้าที่เป็นโหนดราก ถ้าคุณลักษณะใดที่ให้ข้อมูลผลกำไรสูงที่สุด แสดงว่าคุณลักษณะนั้นสามารถจำแนกกลุ่มของข้อมูลได้ดีที่สุด การใช้ค่าข้อมูลผลกำไรจะช่วยลดจำนวนครั้งของการทดสอบในการแยกข้อมูล อีกทั้งยังรับประกันว่าต้นไม้ตัดสินใจที่ได้ไม่มีความซับซ้อนมากจนเกินความจำเป็น ซึ่งค่าข้อมูลผลกำไรนั้นสามารถคำนวณได้จากสมการ 2-4

$$I(S_1, S_2, ..., S_n) = -\sum_{i=1}^{n} \frac{S_i}{S} \log_2 \frac{S_i}{S} \qquad (2\text{-}4)$$

เมื่อ S แทนเซตของข้อมูลซึ่งประกอบด้วยข้อมูล S ระเบียน (Record)

N แทนจำนวนกลุ่มทั้งหมดที่ต่างกันของข้อมูลชุดนั้น

S_i แทนจำนวนข้อมูลที่เป็นสมาชิกของ S และอยู่ในกลุ่มของ C_i

C_i แทนกลุ่มในลำดับที่ i โดยที่ i มีค่าระหว่าง 1 ถึง n

ค่าเอ็นโทรปี (Entropy) ของคุณลักษณะ A(E(A)) ซึ่งมีค่าของคุณลักษณะเป็น $(a_1, a_2, ..., a_n)$ หาได้จากสมการ 2-5

$$E(A) = \sum_{j=1}^{v} \frac{S_{1j} + ... + S_{nj}}{S} I(S_{1j}, S_{2j}, ...S_{nj}) \qquad (2\text{-}5)$$

S_{ij} แทนข้อมูลจำนวนที่เป็นสมาชิกของ S ในกลุ่ม C_i จากการแบ่งข้อมูลด้วยค่าที่เป็นไปได้ของคุณลักษณะ A

ดังนั้นจะสามารถพิจารณาค่าเกณฑ์ผลกำไรได้จากสมการ 2-6

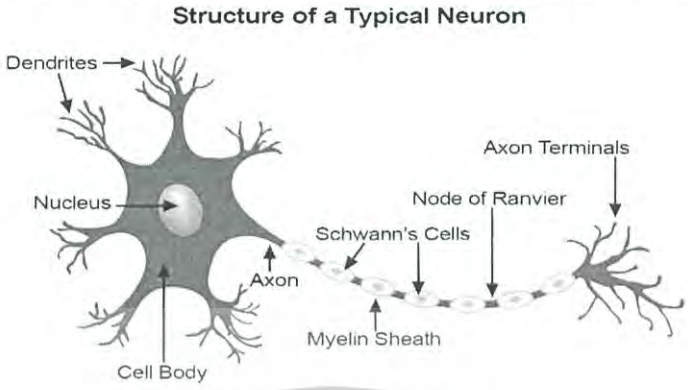
$$Gain(A) = I(S_{1j}, S_{2j}, ..., S_{nj}) - E(A) \qquad (2\text{-}6)$$

2.5 วิธีโครงข่ายประสาทเทียม (Artificial Neural Network)

โครงข่ายประสาทเทียมเป็นศาสตร์ที่จำลองแบบความสามารถของมนุษย์ด้านการเรียนรู้จดจำและจำแนกสิ่งต่างๆ ซึ่งใช้สมองเป็นส่วนสำคัญ ในการประมวลระบบของโครงข่ายประสาทเทียมนั้นจะเลียนแบบการทำงานของระบบสมอง (ศุภกฤษ์ ชูธงชัย, 2546) คือมีการส่งผ่านข้อมูลระหว่างกันโดยมีการเชื่อมต่อของเซลล์ประสาท (Neuron) กันเป็นโครงข่ายร่างแหจำนวนมาก และมีการประมวลผลในลักษณะขนาน (Parallel processing) สาเหตุหลักที่โครงข่ายประสาทเทียมเป็นที่นิยมกันมากขึ้นเนื่องจากมีความยืดหยุ่นในการทำงานสูงและสามารถปรับตัวเองให้ทำงานในสภาพที่เปลี่ยนแปลง อีกทั้งไม่จำเป็นต้องทราบตัวแบบทางคณิตศาสตร์ (Mathematical model) ที่แน่นอนของกระบวนการ เพียงแต่ใช้ชุดข้อมูลที่ประกอบด้วยข้อมูลนำเข้า (Input Data) และข้อมูลเป้าหมาย (Target Data) ของกระบวนการในจำนวนมากพอที่ใช้ในการสอน (Training) โครงข่ายประสาทเทียม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.5.1 ความรู้พื้นฐานของระบบประสาท (Neural System Knowledge)



รูปที่ 2-4 โครงข่ายของเซลล์ประสาท

ภายในสมองของมนุษย์ประกอบด้วยหน่วยประมวลผลขนาดเล็กที่เรียกว่า เซลล์ประสาท (Neurons) ซึ่งจะมีประมาณ 10 เซลล์ ในเซลล์ประสาทแต่ละเซลล์ดังแสดงในรูปที่ 2-4 ประกอบด้วย ใยประสาท (Dendrites) ตัวเซลล์ (Cell body) และเส้นใยประสาท (Axon) ซึ่งแบ่งออกเป็น 4 บริเวณด้วยกัน คือ

1) บริเวณนำกระแสประสาทเข้า (Input region) เป็นบริเวณที่จะมีการนำกระแสประสาท (Nerve impluse) จากเซลล์ประสาทอื่นเข้ามาภายในตัวเซลล์โดยผ่านทางใยประสาทซึ่งมีลักษณะแตกเป็นกิ่งก้านคล้ายต้นไม้และมีจำนวนตั้งแต่ 1 ใยขึ้นไป

2) บริเวณรวมกระแสประสาท (Integration) เป็นบริเวณที่มีการรวมกระแสประสาทก่อนที่จะเข้าสู่บริเวณการนำกระแสประสาทรวมออกจากเซลล์

3) บริเวณการนำกระแสประสาทรวมออกจากเซลล์ (Conduction region) เป็นบริเวณที่จะนำกระแสประสาทรวมออกจากเซลล์ โดยใช้เส้นใยประสาทเป็นทางผ่าน ซึ่งมีเพียง 1 เส้นใยต่อเส้นเดียวเท่านั้น

4) บริเวณนำกระแสประสาทออก (Output region) เป็นบริเวณส่วนปลายของเส้นใยประสาทที่มีการแตกแขนง ใช้ในการถ่ายทอดกระแสประสาทข้ามเซลล์ไปยังเซลล์ประสาทอื่น โดยผ่านทางใยประสาทของเซลล์ประสาทนั้น

2.5.2 การเรียนรู้ของโครงข่ายประสาทเทียม (Artificial Neural Network Learning)

การเรียนรู้ของโครงข่ายประสาทเทียมจะมีประสิทธิภาพเพียงใดนั้นขึ้นอยู่กับค่าถ่วงน้ำหนัก (Weight) ของโครงข่ายที่ทำการออกแบบซึ่งการฝึกหัด (Training) โครงข่ายคือการหาค่าถ่วงน้ำหนักที่เหมาะสมให้กับโครงข่ายนั้นๆ โดยทั่วไปสามารถจำแนกวิธีการเรียนรู้ของโครงข่ายประสาทประสาทเทียมได้เป็น 2 ประเภทคือ การเรียนรู้แบบมีผู้สอนและการเรียนรู้แบบไม่มีผู้สอน

1) การเรียนรู้แบบมีผู้สอน (Supervised learning)

การเรียนรู้แบบมีผู้สอนจะกำหนดข้อมูลฝึกหัด (Training data set) ให้กับโครงข่าย ซึ่งกลุ่มนี้ประกอบด้วยข้อมูลนำเข้าและข้อมูลเป้าหมายที่ต้องการ จากนั้นโครงข่ายจะทำการคำนวณค่าถ่วง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

น้ำหนักที่เหมาะสมให้กับข้อมูลฝึกหัด โดยคำตอบที่ได้จากโครงข่ายจะถูกคำนวณค่าความผิดพลาด (Error Value) ว่ามีความห่างจากคำตอบที่ต้องการของข้อมูลนำเข้าในชุดเดียวกันมากน้อยเพียงใด ถ้ายังมีความผิดพลาดสูงอยู่ การฝึกหัดจะดำเนินต่อจนกว่าค่าความผิดพลาดจะลดลงต่ำกว่าค่าที่ยอมรับได้ (Accepted level) จึงจะหยุดฝึกหัด สุดท้ายค่าถ่วงน้ำหนัก (Weight) ที่ได้จะเป็นเหมือนฟังก์ชันที่ใช้ในการแปลงข้อมูล

2) การเรียนรู้แบบไม่มีผู้สอน (Unsupervised learning)

การเรียนรู้แบบไม่มีผู้สอนจะอาศัยชุดข้อมูลนำเข้าเพียงอย่างเดียวในการฝึกหัดโครงข่าย โดยไม่มีข้อมูลเป้าหมาย แต่จะใช้ข้อมูลนำออก (Output data) จากโครงข่ายแทน เมื่อป้อนข้อมูลเข้าโดยอาศัยค่าถ่วงน้ำหนักเป็นตัวแยกความแตกต่างของข้อมูลนำเข้าและนำไปเก็บไว้ในโหนดข้อมูลนำออกของโครงข่าย ซึ่งวัตถุประสงค์เพื่อใช้ในการจำแนกชุดข้อมูล

2.5.3 การเชื่อมโยงของโครงข่ายประสาทเทียม (Artificial Neural Network Linking)

เพื่อให้โครงข่ายประสาทเทียมสามารถเรียนรู้ได้อย่างมีประสิทธิภาพ จำเป็นต้องมีการเชื่อมโยงกันระหว่างเซลล์ประสาท โดยทั่วไปสามารถแบ่งการเชื่อมโยงของโครงข่ายได้ 2 ลักษณะ คือ

1) โครงข่ายแบบส่งสัญญาณไปข้างหน้า (Feed-forward network)

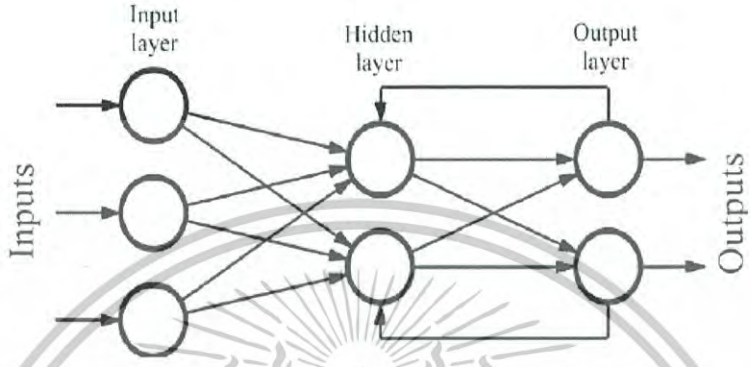
เป็นโครงข่ายที่การประมวลผลจะอาศัยชุดข้อมูลปัจจุบันและส่งค่าที่ประมวลผลได้ไปยังชั้นถัดๆไป กล่าวคือ โครงข่ายชนิดนี้จะประกอบด้วยชั้นต่างๆ โดยชั้นแรกจะเป็นชั้นนำเข้า (Input Layer) และชั้นสุดท้ายเป็นชั้นนำออก (Output Layer) ส่วนระหว่างชั้นนำเข้ากับชั้นนำออกอาจมีหรือไม่มีชั้นซ่อน (Hidden Layers) อยู่ภายใน ซึ่งขึ้นอยู่กับกฎการเรียนรู้ (Learning Rule) ที่ใช้ในการสอนโครงข่าย เช่น ถ้าเป็นโครงข่ายเพอร์เซปตรอนแบบหลายชั้น (Multi-layer Perceptron) จะมีชั้นซ่อนอยู่ระหว่างชั้นนำเข้ากับชั้นนำออก ซึ่งอาจมีมากกว่า 1 ชั้นได้ การเชื่อมต่อระหว่างชั้นโครงข่ายแบบส่งสัญญาณไปข้างหน้าจะมีค่าถ่วงน้ำหนักเป็นตัวเชื่อมและสัญญาณนำเข้าที่เข้ามาจะถูกส่งไปตามทิศทางของลูกศรจนถึงชั้นนำออกโดยไม่มีการป้อนกลับ สามารถแสดงตัวแบบได้ดังรูปที่ 2-5

รูปที่ 2-5 ลักษณะโครงข่ายประสาทเทียมแบบส่งสัญญาณไปข้างหน้า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2) โครงข่ายแบบมีการย้อนกลับ (Feed-back Network)

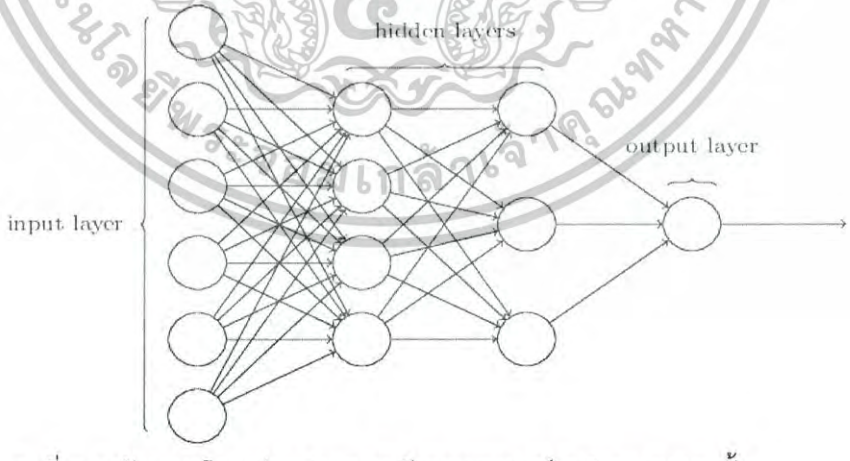
โครงข่ายชนิดนี้มีชื่อเรียกอีกชื่อหนึ่งว่า โครงข่ายหันกลับ (Recurrent Network) เป็นโครงข่ายที่จะอาศัยข้อมูลทั้งข้อมูลในปัจจุบันและข้อมูลที่มีการถ่วงเวลามาใช้ในการประมวลผลของโครงข่ายประสาทเทียม สามารถแสดงตัวแบบโครงข่ายที่มีการย้อนได้ดังรูปที่ 2-6



รูปที่ 2-6 ลักษณะโครงข่ายประสาทเทียมแบบย้อนกลับ

2.5.4 การแพร่แบบย้อนกลับ (Back-propagation)

การแพร่แบบย้อนกลับเป็นขั้นตอนที่ใช้สอนโครงข่ายประสาทเทียมแบบเพอร์เซปตรอนหลายชั้น ซึ่งแบบจำลองโครงข่ายประสาทเทียมมีการเชื่อมโยงกันเป็นโครงข่ายแบบเป็นชั้นๆ โครงข่ายชนิดนี้มีการเชื่อมโยงกัน 3 ชั้น ประกอบด้วย ชั้นนำเข้า (Input Layer) ถัดมาเป็นชั้นซ่อน (Hidden Layer) และชั้นสุดท้ายคือชั้นนำออก (Output Layer) ดังรูปที่ 2-7



รูปที่ 2-7 ลักษณะโครงข่ายประสาทเทียมแบบเพอร์เซปตรอนหลายชั้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.5.5 ปัจจัยที่ส่งผลต่อการเรียนรู้การแพร่แบบย้อนกลับ

1) การกำหนดค่าเริ่มต้นของค่าถ่วงน้ำหนัก

ก่อนที่จะทำการสอนโครงข่ายประสาทเทียมแบบหลายชั้น จำเป็นต้องกำหนดค่าเริ่มต้นให้ค่าถ่วงน้ำหนักที่เชื่อมโยงระหว่างชั้นทุกชั้น โดยค่านี้จะเป็นเลขจำนวนจริงที่มีค่าน้อยๆ ที่ได้มาจากการสุ่มค่าเริ่มต้น (Randomness)

2) การกำหนดเกณฑ์การหยุดฝึกหัด

เกณฑ์ในการหยุดฝึกหัดนั้นขึ้นอยู่กับผู้ทำการออกแบบโครงข่ายประสาทเทียมว่าต้องการที่จะให้โครงข่ายประสาทเทียมมีความแม่นยำเพียงใด โดยทั่วไปนิยมใช้ค่าดัชนีที่ชี้ถ่วงค่าความผิดพลาดของระบบได้ ในงานวิจัยนี้ใช้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error : MSE)

3) อัตราการเรียนรู้ (Learning Rate, η)

อัตราการเรียนรู้เป็นค่าสัมประสิทธิ์ที่แสดงถึงการเรียนรู้ของโครงข่าย โดยทั่วไปค่าที่เหมาะสมจะอยู่ในช่วง 0.05 ถึง 0.5 ถ้าอัตราการเรียนรู้มีค่าสูงแสดงว่ากำหนดให้โครงข่ายมีการเปลี่ยนแปลงค่าถ่วงน้ำหนักที่มาก ในทางตรงกันข้ามถ้าอัตราการเรียนรู้มีค่าต่ำแสดงว่ากำหนดให้โครงข่ายมีการเปลี่ยนแปลงค่าถ่วงน้ำหนักที่น้อย ซึ่งจำเป็นต้องใช้เวลาในการเรียนรู้ที่มากขึ้น แต่จะมีข้อดีคือโครงข่ายจะมีเสถียรภาพและไม่เกิดการแกว่ง (Oscillation) ขณะที่ทำการเรียนรู้

4) ค่าคงที่โมเมนตัม (Momentum Constant, α)

ค่าคงที่โมเมนตัมเป็นค่าสัมประสิทธิ์ที่ช่วยหน่วงไม่ให้การเปลี่ยนแปลงค่าถ่วงน้ำหนักนั้นมีค่ามากเกินไป เป็นการเพิ่มเสถียรภาพให้กับโครงข่ายได้อีกทางหนึ่ง ซึ่งค่าโมเมนตัมที่เหมาะสมจะมีค่าเข้าใกล้ 1.0 และควรจะกำหนดให้สอดคล้องกับอัตราการเรียนรู้ด้วย เช่น ถ้าอัตราการเรียนรู้สูงก็ควรที่จะมีค่าโมเมนตัมที่ต่ำ ทำให้การเปลี่ยนแปลงค่าถ่วงน้ำหนักไม่มากจนเกินไป

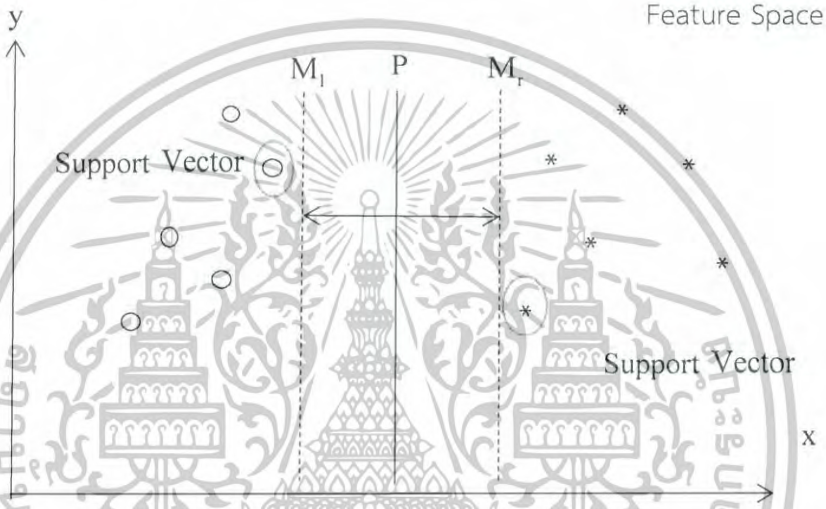
การประยุกต์ใช้งานโครงข่ายประสาทเทียม เนื่องจากความสามารถในการจำลองพฤติกรรมทางกายภาพของระบบที่มีความซับซ้อนจากข้อมูลที่ป้อนให้เรียนรู้ การประยุกต์ใช้ข่ายงานระบบประสาทจึงเป็นทางเลือกใหม่ในการควบคุม ซึ่งมีผู้นำมาประยุกต์ใช้งานหลายประเภท ได้แก่ งานการจดจำรูปแบบที่มีความไม่แน่นอน เช่น ลายมือ ลายเซ็น ตัวอักษร รูปหน้า งานการประมาณค่าฟังก์ชันหรือการประมาณความสัมพันธ์ (มี inputs และ outputs แต่ไม่ทราบว่า inputs กับ outputs มีความสัมพันธ์กันอย่างไร) งานที่สิ่งแวดล้อมเปลี่ยนแปลงอยู่เสมอ (โครงข่ายประสาทเทียมสามารถปรับตัวเองได้) งานจัดหมวดหมู่และแยกแยะสิ่งของ งานทำนาย เช่น พยากรณ์อากาศ พยากรณ์หุ้น การประยุกต์ใช้โครงข่ายประสาทเทียมควบคุมกระบวนการทางเคมีโดยวิธีพยากรณ์แบบจำลอง (Model Predictive Control) การประยุกต์ใช้โครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับในการทำนายพลังงานความร้อนที่สะสมอยู่ในตัวอาคาร การใช้โครงข่ายประสาทเทียมในการหาไซโครเมตริกชาร์ทและการประยุกต์ใช้ข่ายงานระบบประสาทควบคุมระบบ HVAC

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.6 วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)

2.6.1 แนวความคิดของซัพพอร์ตเวกเตอร์แมชชีน

ซัพพอร์ตเวกเตอร์แมชชีนเป็นสมการที่ใช้ในการจำแนกค่าคุณลักษณะของ 2 กลุ่มที่วางตัวอยู่ในพื้นที่คุณลักษณะ (Feature Space) ออกจากกันโดยจะสร้างเส้นแบ่ง (Plane) ที่เป็นเส้นตรงขึ้นมา และเพื่อให้ทราบว่าเส้นตรงที่แบ่ง 2 กลุ่มออกจากกันนั้น เส้นตรงใดที่เป็นเส้นที่ดีที่สุด โดยเส้นตรงนั้นจะเพิ่มเส้นขอบ (Margin) ออกไปทั้งสองข้าง โดยเส้นขอบที่เพิ่มนั้นจะขนานกับเส้นเดิมเสมอ เส้นขอบที่เพิ่มขึ้นมานี้จะขยายออกไปจนกว่าจะสัมผัสกับค่าของกลุ่มตัวอย่างที่ใกล้ที่สุด (จิรา แก้วสุวรรณ์, 2549) ดังรูปที่ 2-8



รูปที่ 2-8 การขยายตัวของเส้นขอบ

จากรูปที่ 2-8 เส้น M_l และ M_r คือเส้นขอบที่ขยายออกไปด้านซ้ายและขวาตามลำดับ และ P คือเส้นแบ่งข้อมูลทั้ง 2 กลุ่ม เมื่อเส้น M_l และ M_r ขยายออกไปจนสัมผัสค่าข้อมูลที่ใกล้ที่สุด ซึ่งข้อมูลที่อยู่บนเส้นขอบของทั้งสองฝั่งนั้นเรียกว่า ซัพพอร์ตเวกเตอร์ (Support Vector) จะวัดค่าระยะความห่างของเส้นขอบ โดยเส้น P จะเปลี่ยนความชันไปเรื่อยๆ เพื่อที่จะหาความกว้างสูงสุดของเส้นขอบ

กระบวนการโดยรวมของซัพพอร์ตเวกเตอร์แมชชีนนั้น เป็นการหาค่าความชันของเส้น P ที่มีขนาดของเส้นขอบสูงสุด

2.6.2 สมการพื้นฐานของซัพพอร์ตเวกเตอร์แมชชีน

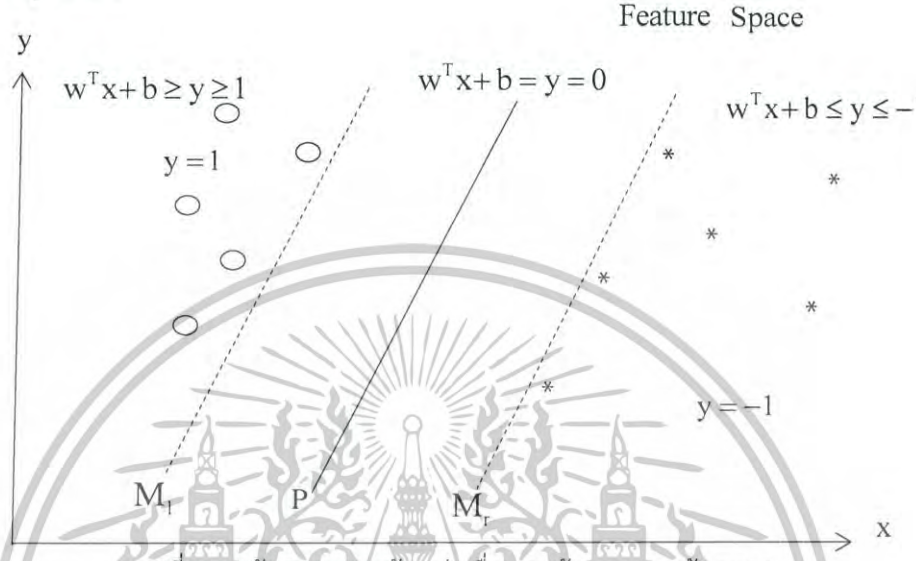
ถ้านำแนวคิดของซัพพอร์ตเวกเตอร์แมชชีนที่กล่าวไปแล้วในข้อ 2.6.1 มาเขียนเป็นสมการเพื่อใช้ในการแก้ปัญหา โดยข้อมูลที่นำมาวางลงในพื้นที่คุณลักษณะนั้นเป็นกลุ่มข้อมูลที่อยู่ในรูปของเวกเตอร์ สมการ 2-7

$$X = ((x_1, y_1), ..., (x_i, y_i)) \quad (2\text{-}7)$$

เมื่อ X คือ ชุดค่าคุณลักษณะ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ค่าคุณลักษณะที่วางอยู่ในพื้นที่คุณลักษณะจะถูกแบ่งด้วยเส้นตรงดังรูปที่ 2-9 และเมื่อนำเส้นตรงมาแทนค่าด้วยสมการเส้นตรง $y = mx + b$ โดยมีการกำหนดกลุ่มของข้อมูลทั้งสองฝั่งเป็นเพียงสองค่าที่ซึ่งแทนด้วยค่า y เพื่อให้ข้อมูลที่อยู่ในกลุ่มเดียวกันที่มาจากหลายค่ากลายเป็นค่าเดียว ดังสมการในรูปที่ 2-9



รูปที่ 2-9 เส้นขอบและเส้นแบ่งเมื่อแทนด้วยสมการเส้นตรง

จากรูปที่ 2-9 เส้นตรง M_l แทนด้วยสมการ $w^T x+b \geq y \geq 1$ ซึ่งข้อมูล y ที่มากกว่า 1 ก็จะถูกกำหนดค่าใหม่โดยให้ y เท่ากับ 1 และพจน์ w ก็คือค่าความชัน เช่นเดียวกับเส้นตรง M_r ที่ค่าของ y จะถูกกำหนดค่าใหม่เมื่อ y ที่น้อยกว่า -1 ให้เท่ากับ -1 ดังนั้นสมการที่กำหนดขึ้นใหม่จากสมการเส้นขอบ 2-8 และ 2-9 สามารถกำหนดได้ดังสมการ 2-10

เมื่อ $w^T x+b \geq y$ กำหนด $y = 1$ (2-8)

$w^T x+b \leq y$ กำหนด $y = -1$ (2-9)

$y(w^T x+b)-1 \geq 0$ (2-10)

โดย y คือ ค่ากลุ่มข้อมูล (1,-1)

w คือ ค่าความชัน

x คือ ค่าคุณลักษณะ

b คือ ค่าคงที่ (ค่าตัดแกน y)

2.6.3 ค่าความกว้างเส้นขอบ (Margin)

การคำนวณความกว้างของเส้นขอบต้องทำการคำนวณพจน์ w ให้อยู่ในรูปปกติมาตรฐาน (Normalization) โดยคำนวณจากสมการที่ 2-8 และ 2-9 เมื่อแทนค่า y ลงไปแล้ว

$$w^T x^+ + b = 1$$

$$w^T x^- + b = -1$$

$$w^T (x^+ - x^-) = 2 \quad (2\text{-}11)$$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากสูตร $$M = (\frac{w}{\|w\|})^T(x^+ - x^-) \quad (2\text{-}12)$$

นำสมการ 2-11 แทนลงในสมการ 2-12 จะได้

$$M = \frac{2}{\|w\|} \quad (2\text{-}13)$$

โดยที่ M คือ ความกว้างของเส้นขอบ (Margin)

2.6.4 การแก้ปัญหา Dual Problem

หลังจากที่ได้สมการที่ 2-10 และ 2-13 ของการหาเส้นแบ่งและค่าความกว้างตามลำดับแล้ว จึงทำการแก้สมการนั้นต่อด้วย Lagrangian Dual Problem เพื่อหาค่าในพจน์ของ w ในสมการ 2-10

$$L(w,b,\alpha) = \frac{1}{2}(w,w) - \sum_{i=1}^{N}\alpha_i[y_i((w_i,w_i)+b)-1] \quad (2\text{-}14)$$

เมื่อ $\alpha_i \geq 0;\ i = 1,\ldots\ldots, N$

สมการที่ 2-14 นำมาหาอนุพันธ์ (Differential) จะได้

$$\frac{\partial L(w,b,\alpha)}{\partial w} = w - \sum_{i=1}^{N}(\alpha_i y_i x_i) = 0 \quad (2\text{-}15)$$

$$\frac{\partial L(w,b,\alpha)}{\partial b} = \sum_{i=1}^{N}\alpha_i y_i = 0 \quad (2\text{-}16)$$

หลังจากได้สมการที่ผ่านแก้ปัญหาด้วย Dual Problem แล้วค่า w จะลดรูป และหาได้จาก

$$w = \sum_{i=1}^{N}(\alpha_i y_i x_i) \quad (2\text{-}17)$$

เมื่อ α คือ สัมประสิทธิ์คงที่

เมื่อนำค่า w ไปใส่ในสมการที่ 2-10 ซึ่งเป็นสมการในการหาเส้นแบ่ง จะได้

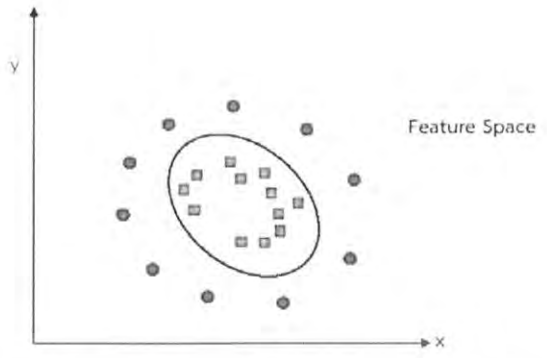
$$y_i(\sum_{i=1}^{N}\alpha_i y_i(x_i^T x)+b)-1 \geq 0 \quad (2\text{-}18)$$

จากสมการที่ 2-15 คือการหาค่าคงที่ของสัมประสิทธิ์ α เพื่อนำมาใช้บอกการวางตัวของเส้นแบ่งข้อมูลต่อไป

2.6.5 เคอร์เนล (Kernel)

ในโลกความเป็นจริงนั้นข้อมูล 2 กลุ่มไม่ได้วางตัวในพื้นที่คุณลักษณะ และไม่สามารถแบ่งได้โดยเส้นตรง แต่ข้อมูลนั้นอาจจะจับกลุ่มกันในตำแหน่งต่างๆ ดังนั้นจึงเป็นปัญหาทำให้ไม่สามารถที่จะใช้สมการซัพพอร์ตเวกเตอร์แมชชีนแบบเชิงเส้นได้ ดังนั้นจะต้องมีเครื่องมือมาช่วยให้ข้อมูลเหล่านั้นเรียงตัวใหม่ในพื้นที่ เรียกว่า พื้นที่หลายมิติ (Higher Dimensional Space)

เอกสารนี้สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 2-10 รูปแบบการวางตัวที่ไม่สามารถแบ่งด้วยเส้นตรงได้

ในเคอร์เนลนั้นคือการคูณกันของชุดเวกเตอร์ของ x ใดๆ

$$K(x_i, x_j) = x_i^T x_j \quad (2\text{-}19)$$

เคอร์เนลที่นิยมใช้มีอยู่ 3 ชนิด คือ

1) โพลิโนเมียล (Polynomial)

$$K(x_i, x_j) = (x_i x_j + 1)^d \quad (2\text{-}20)$$

เมื่อ d คือค่าเลขยกกำลัง

2) ฟังก์ชันเบสิสรัศมี (Radial Basis Funtion : RBF)

$$K(x_i, x_j) = e^{\frac{-|x_i - x_j|^2}{2\sigma^2}} \quad (2\text{-}21)$$

เมื่อ σ คือ ค่าพารามิเตอร์

3) ซิกมอยด์ (Sigmoid)

$$K(x_i, x_j) = \tanh(kx_i x_j + \mu) \quad (2\text{-}22)$$

เมื่อ k, μ คือค่าพารามิเตอร์

ดังนั้นจากสมการของเคอร์เนลนั้นสามารถที่จะแทนลงไปในตำแหน่งของ $x_i^T x_j$ ในสมการที่ 2-18 จึงเขียนเป็นสมการใหม่ดังนี้

$$y_i(\sum_{i=1}^{N} \alpha_i y_i K(x_i x_j) + b) - 1 \geq 0 \quad (2\text{-}23)$$

สมการที่ 2-23 เป็นสมการที่ใช้ในขั้นตอนที่จะเรียนรู้ว่าจะวางตำแหน่งเส้นแบ่งไว้ที่ตำแหน่งใดโดยทำงานร่วมกับเคอร์เนล เพื่อแปลงให้ข้อมูลที่ยากต่อการแบ่งแบบเชิงเส้นสามารถแบ่งได้เมื่อทำให้เป็นข้อมูลแบบหลายมิติ (Higher Dimension) ดังนั้นจึงมีอีกสมการหนึ่งที่ใช้ค่า w และ b เดิมมาจัดตำแหน่งของข้อมูลเพื่อที่ให้ทราบว่าข้อมูลนั้นเป็นกลุ่มใด กำหนดได้ดังสมการที่ 2-24

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{N} \alpha_i y_i K(x_i x_j) + b) \quad (2\text{-}24)$$

เมื่อ f(x) คือค่า y หาในรูปของ x

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.7 การเปรียบเทียบประสิทธิภาพของวิธีการจำแนกกลุ่ม

2.7.1 เมตริกซ์ความสับสน (Confusion Matrix) เป็นรูปแบบตารางที่เฉพาะเจาะจงที่นำผลลัพธ์จากการทำนายมาใส่ในรูปตารางเมตริกซ์ ซึ่งจะช่วยให้ง่ายต่อการมองเห็นค่าทำนายของอัลกอริทึมดังรูปที่ 2-11

ชั้นของค่าจริง	ชั้นของค่าทำนาย	
	A	B
A	TP	FN
B	FP	TN

รูปที่ 2-11 เมตริกซ์ความสับสน (Confusion Matrix)

โดยที่ True Positive (TP) คือจำนวนข้อมูลที่จำแนกถูกว่าเป็นชั้น A

True Negative (TN) คือจำนวนข้อมูลที่จำแนกถูกว่าเป็นชั้น B

False Positive (FP) คือจำนวนข้อมูลที่จำแนกผิดว่าเป็นชั้น A ซึ่งชั้นที่แท้จริงเป็นชั้น B

False Negative (FN) คือจำนวนข้อมูลที่จำแนกผิดว่าเป็นชั้น B ซึ่งชั้นที่แท้จริงเป็นชั้น A

ค่าการทำนาย (Prediction) คือส่วนที่แสดงผลการทำนายของแต่ละตัวอย่างของแต่ละชุดข้อมูล

ค่าการจำแนกได้ถูกต้อง (Correctly Classified Instances) คือค่าที่บอกว่าชุดข้อมูลมีอัตราการทำนายถูกต้องและผิดพลาดเท่าไร

2.7.2 ค่าความถูกต้อง (Accuracy) คือการแสดงการวัดที่ได้มีความถูกต้องในรูปอัตราส่วน

$$\text{Accuracy} = \frac{\text{จำนวนข้อมูลที่จำแนกถูกว่าเป็นชั้น A และ B}}{\text{จำนวนข้อมูลทั้งหมด}}$$

$$= \frac{TP+TN}{TP+TN+FP+FN}$$

2.7.3 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error : MSE) เป็นมาตรวัดการประเมินค่าได้ดี เนื่องจากความคลาดเคลื่อนกำลังสองเฉลี่ยประกอบด้วยทั้งความเอนเอียงและความแปรปรวน (สายชล สินสมบูรณ์ทอง, 2558)

$$\text{Mean Square Error}(\hat{\theta}) = \text{Var}(\hat{\theta}) + (\text{Bias}(\hat{\theta}))^2$$

$$\text{Mean Square Error} = \frac{\sum_{i=1}^{n}\left(y_i - \widehat{y_i}\right)^2}{n}$$

โดยที่ y_i แทน ค่าจริง

$\widehat{y_i}$ แทน ค่าพยากรณ์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.7.4 ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย (Mean Absolute Deviation : MAD หรือ Mean Absolute Error : MAE) คือ ค่าวัดความถูกต้องของการพยากรณ์ที่วัดจากค่าความคลาดเคลื่อนโดยไม่คำนึงถึงทิศทางของความคลาดเคลื่อน MAD มีหน่วยวัดหน่วยเดียวกับค่าสังเกต (สายชล สินสมบูรณ์ทอง, 2558)

$$\text{Mean Absolute Deviation} = \frac{\sum_{i=1}^{n}\left|y_i - \widehat{y_i}\right|}{n}$$

โดยที่ y_i แทน ค่าจริง $\widehat{y_i}$ แทน ค่าพยากรณ์

2.8 งานวิจัยที่เกี่ยวข้อง

นิเวศ จิระวิชิตชัย (2553) ศึกษาเรื่องการค้นหาเทคนิคเหมืองข้อมูลเพื่อสร้างโมเดลการวิเคราะห์โรคอัตโนมัติ โดยมีวัตถุประสงค์เพื่อค้นหาเทคนิคด้านเหมืองข้อมูล เพื่อสร้างโมเดลการวิเคราะห์โรคอัตโนมัติทดสอบประสิทธิภาพในการจำแนกกลุ่ม (Classification) สำหรับข้อมูลทางการแพทย์ ผลจากการวิจัยพบว่าทุกโมเดลที่สร้างขึ้นมีประสิทธิภาพในการจำแนกกลุ่มของโรคในระดับ 80 % ขึ้นไป เมื่อไม่ลดคุณลักษณะ และเมื่อเรียงค่าความถูกต้อง (Accuracy) แยกตามประเภทของข้อมูลพบว่า กลุ่มข้อมูล Hypothyroid การสร้างโมเดลด้วยอัลกอริทึม Decision Tree ให้ประสิทธิภาพดีที่สุด 99.57% กลุ่มข้อมูล Leukemia การสร้างโมเดลด้วยอัลกอริทึม Naïve Bayes กับ Support Vector Machine ให้ประสิทธิภาพดีที่สุด 98.61% กลุ่มข้อมูล Breast-w การสร้างโมเดลด้วยอัลกอริทึม Support Vector Machine ให้ประสิทธิภาพดีที่สุด 96.99% กลุ่มข้อมูล Lymphography การสร้างโมเดลด้วยอัลกอริทึม Support Vector Machine ให้ประสิทธิภาพดีที่สุด 86.48% กลุ่มข้อมูล Hepatitis การสร้างโมเดลด้วยอัลกอริทึม Radial Basis Function กับ K-Nearest Neighbor ให้ประสิทธิภาพดีที่สุด 85.80% กลุ่มข้อมูล Heart-c การสร้างโมเดลด้วยอัลกอริทึม Support Vector Machine ให้ประสิทธิภาพดีที่สุด 84.15% กลุ่มข้อมูล Heart-statlog การสร้างโมเดลด้วยอัลกอริทึม Radial Basis Function กับ Support Vector Machine ให้ประสิทธิภาพดีที่สุด 84.07% ตามลำดับ ซึ่งการสร้างโมเดลดังกล่าวมีคุณภาพในระดับที่ยอมรับได้และสามารถนำไปพัฒนาเป็นซอฟต์แวร์ในการวินิจฉัยโรคอัตโนมัติได้

กิตติพล วิแสง และคณะ (2552) ศึกษาเรื่องการวิเคราะห์ปัจจัยเสี่ยงของโรคเบาหวาน พบว่าปัจจัยเสี่ยงของโรคเบาหวานมีความสำคัญอย่างยิ่งในการนำมาเป็นเครื่องมือประเมินการเกิดโรคเบาหวาน แทนการตรวจเลือด บทความนี้ศึกษาและวิเคราะห์ปัจจัยเสี่ยงและลำดับของวิจัยเสี่ยงในการเกิดโรคเบาหวาน โดยตัวแบบที่ใช้ศึกษาคือ ตัวแบบ Backpropagation Neural Networks, Radial Basis Function Network และตัวแบบ Naïve Bayes ผลการวินิจฉัยทางการแพทย์ได้ถูกนำมาอ้างอิงเพื่อวัดความถูกต้องของตัวแบบ ผลการศึกษาแสดงให้เห็นว่าตัวแบบ Back-propagation Network ให้ผลลัพธ์ที่มีความถูกต้องมากที่สุด ปัจจัยเสี่ยงและลำดับของปัจจัยเสี่ยงสามารถตัดสินได้จากตัวแบบที่ศึกษา ซึ่งข้อมูลเหล่านี้สามารถนำไปสร้างเป็นแบบประเมินความเสี่ยงโรคเบาหวานโดย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ไม่ต้องอาศัยการตรวจเลือดได้ รวมทั้งเป็นเครื่องมือในการประเมินตนเองและสามารถพัฒนาเป็นระบบผู้เชี่ยวชาญสำหรับการวินิจฉัยโรคเบาหวานต่อไปได้

เดช ธรรมศิริ วาทินี นุ้ยเพียร ภัทราวุฒิ แสงศิริ ภรัณยา อำมฤครัตน์ ณรงค์ โพธิและพยุง มีสัจ (2552) ศึกษาเรื่องการให้คะแนนสินเชื่อโดยวิธีการทำเหมืองข้อมูลด้วยเทคนิคซับพอร์ตเวกเตอร์แมชชีนรวมทั้งการเลือกใช้ลักษณะที่เหมาะสมร่วมกับการหาค่าพารามิเตอร์ที่เหมาะสมด้วยวิธีค้นหาแบบกริช โดยมีวัตถุประสงค์เพื่อสร้างโมเดลที่มีประสิทธิภาพ สำหรับการจำแนกข้อมูลเครดิต โดยนำข้อมูลจาก UCI และข้อมูล Bankruptcy dataset มาเป็นข้อมูลใช้สำหรับทดสอบ ผลที่ได้แสดงให้เห็นว่าการนำเอาเทคนิคซัพพอร์ต-เวกเตอร์แมชชีนสำหรับการจำแนกข้อมูลร่วมกับวิธีการคัดเลือกคุณลักษณะที่เหมาะสม (Feature Selection) อีกทั้งการหาค่าพารามิเตอร์ที่เหมาะสมที่สุดสำหรับซัพพอร์ตเวกเตอร์แมชชีน โดยใช้วิธีการค้นหาแบบกริช ในการวิจัยพบว่ามีความแม่นยำมากกว่าวิธีการต้นไม้ตัดสินใจ ที่มีความแม่นยำโดยเฉลี่ย 79.48% วิธีโครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับที่มีความแม่นยำโดยเฉลี่ย 82.84 % รวมทั้งวิธีซัพพอร์ตเวกเตอร์แมชชีนเพียงอย่างเดียวให้ความแม่นยำโดยเฉลี่ยที่ 83.96% โดยการใช้ t-test ร่วมกับวิธีซัพพอร์ตเวกเตอร์แมชชีนและวิธีการค้นหาแบบกริชจะให้ค่าเฉลี่ยความแม่นยำสูงที่สุดถึง 85.68%

ทิพย์ธิดา วงศ์พิพันธ์ (2555) ศึกษาเรื่องการใช้เหมืองข้อมูลช่วยในการตัดสินใจการให้สินเชื่อ กรณีศึกษา: บริษัท กรุงไทยคาร์เร้นท์ แอนด์ ลีส จำกัด (มหาชน) โดยงานวิจัยนี้เป็นการนำเหมืองข้อมูลช่วยในการตัดสินใจในการให้สินเชื่อของ บริษัท กรุงไทยคาร์เร้นท์ แอนด์ ลีส จำกัด (มหาชน) มาเป็นแนวทางสนับสนุนการตัดสินใจการอนุมัติสินเชื่อของบริษัท ซึ่งได้ทดลองจำแนกกลุ่มข้อมูลด้วยเทคนิคการจำแนกกลุ่ม 3 เทคนิค ได้แก่ การจำแนกกลุ่มแบบ Decision Tree:C4.5 หรือเรียกว่าโมเดล Classifires tree J48 โมเดลการจำแนกกลุ่มแบบ Decision Rule:Part หรือเรียกว่าโมเดล Classifires PART และ โมเดลการจำแนกกลุ่มแบบ Bayes: Naïve Bayes หรือเรียกว่าโมเดล Classifires Naïve Bayes ผลลัพธ์ที่ได้คือกฎที่ใช้ในการจำแนกลูกค้ากลุ่มดีและลูกค้ากลุ่มไม่ดี ซึ่งผลการทดลองพบว่า การจำแนกกลุ่มข้อมูลด้วยเทคนิค J48 ให้ผลลัพธ์ดีที่สุด เนื่องจากมีค่าความถูกต้องมากที่สุด และสามารถแสดงกฎให้สามารถนำไปวิเคราะห์กฎต่อไปได้ โดยมีผลลัพธ์ 90.47% และได้กฎจำนวน 40 กฎ

R. Priya, and P. Aruna, (2012) ได้วิจัยเรื่องวิธีซัพพอร์ตเวกเตอร์แมชชีนและวิธีโครงข่ายประสาทในการวินิจฉัยโรคเบาหวานในจอประสาทตา โดยโรคเบาหวานในจอประสาทตาเป็นโรคทางตาที่มีสาเหตุมาจากโรคแทรกซ้อนของโรคเบาหวานและเราควรจะตรวจสอบโรคก่อนเพื่อหาประสิทธิภาพในการรักษา ในขณะที่มีความก้าวหน้าในโรคเบาหวาน สายตาของคนไข้อาจจะเริ่มเสื่อมถอยลงและนำไปสู่โรคเบาหวานในจอประสาทตาได้ ผลการจำแนกมี 2 กลุ่ม คือ ไม่เป็นโรคเบาหวานในจอประสาทตาและเป็นโรคเบาหวานในจอประสาทตา ในงานวิจัยนี้ทำการตรวจสอบโรคเบาหวานในจอประสาทตา 2 วิธีคือ วิธีโครงข่ายประสาทเทียมและวิธีซัพพอร์ตเวกเตอร์แมชชีน ผลการทดลองพบว่าวิธีซัพพอร์ตเวกแมชชีนดีกว่าวิธีโครงข่ายประสาทเทียม โดยที่วิธีซัพพอร์ตเวกเตอร์แมชชีนมีค่าความถูกต้อง 97.61% ส่วนวิธีโครงข่ายประสาทเทียมมีค่าความถูกต้อง 89.60%

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Sriwiboon, N. (2016) ศึกษาเรื่องการเปรียบเทียบประสิทธิภาพอัลกอริธึมเหมืองข้อมูลเพื่อวิเคราะห์ปัจจัยที่ส่งผลต่อการเกิดโรคมะเร็ง ซึ่งโรคมะเร็งเป็นสาเหตุการเสียชีวิตของประชากรไทยเป็นอันดับ 1 ก่อนหน้านี้การวินิจฉัยการเกิดโรคมะเร็งทำได้เช่นการสอบถามประวัติ และการทดสอบเลือดหรือของเหลวภายในร่างกายโดยทดสอบในห้องปฏิบัติการ เป็นต้น ดังนั้นงานวิจัยนี้จึงนำเทคนิคการทำเหมืองข้อมูลมาประยุกต์ใช้กับการตรวจวิเคราะห์การเกิดโรคมะเร็ง โดยเปรียบเทียบอัลกอริธึมการทำเหมืองข้อมูลประกอบด้วย อัลกอริธึม C4.5 อัลกอริธึม k-Nearest Neighbor และอัลกอริธึม Naïve Bayes ผลการเปรียบเทียบพบว่าอัลกอริธึม C4.5 มีประสิทธิภาพสูงสุดที่ 98.63% แล้วนำแบบจำลองที่มีค่าประสิทธิภาพสูงที่สุดเป็นแบบจำลองสำหรับวิเคราะห์ปัจจัยที่ส่งผลต่อการเกิดโรคมะเร็งพบว่าผู้ที่สูบบุหรี่จะมีความเสี่ยงต่อการเป็นโรคมะเร็ง และสามารถนำกฎการจำแนกข้อมูลที่ได้ไปพัฒนาเป็นระบบตรวจวิเคราะห์ปัจจัยที่ส่งผลต่อการเกิดโรคมะเร็งได้

ภรัณยา อามฤครัตนและคณะ (2552) ทำการเปรียบเทียบประสิทธิภาพของแบบจำลองในการคัดเลือกและจำแนกข้อมูลโดยวิธีโครงข่ายประสาทเทียมแบบ Multi-Layer Perceptron (MLP) และซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machines - SVM) การวัดประสิทธิภาพสามารถวัดได้จากความถูกต้องของการจำแนกกลุ่มของข้อมูลโดยนับจากค่าความถูกต้องของการจำแนกประเภทข้อมูล ผลการทดลองที่ได้พบว่าการใช้ SVM ที่ใช้ kernel ด้วย rbf ในการจำแนกข้อมูลนั้นจะมีประสิทธิภาพที่ดีกว่าการใช้แบบจำลองโครงข่ายประสาทเทียม MLP

ชณัฐดาภรณ์ เย็นประเสริฐ (2557) ทำการเพื่อเปรียบเทียบความแม่นยำการพยากรณ์สถานการณ์ชำระหนี้ของลูกหนี้ โดยใช้เทคนิคการถดถอยโลจิสติก เทคนิคเนอีฟเบส์ และเทคนิคต้นไม้ตัดสินใจ ข้อมูลที่ใช้ในการวิจัย คือ ข้อมูลประวัติสถานการณ์ชำระหนี้ของลูกหนี้จากสถาบันการเงินแห่งหนึ่ง จำนวน 392 ข้อมูล ผลลัพธ์จากวิธีการพยากรณ์ความแม่นยำของสถานการณ์ชำระหนี้ของลูกค้าพบว่า เทคนิคต้นไม้ตัดสินใจมีความแม่นยำเท่ากับ 89.80% มีความคลาดเคลื่อนเฉลี่ยสัมพัทธ์ 10.20% รากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 0.30 รองลงมาคือ เทคนิคการถดถอยโลจิสติกมีความแม่นยำเท่ากับ 89.80% ความคลาดเคลื่อนสัมพัทธ์เฉลี่ยเท่ากับ 10.20% รากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 0.31 และเทคนิคนาอีฟเบส์มีความแม่นยำเท่ากับ 88.78% ความคลาดเคลื่อนเฉลี่ยสัมพัทธ์เท่ากับ 11.22% รากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 0.33 จากผลการทดลองกรณีศึกษาพบว่า เทคนิคต้นไม้ตัดสินใจและเทคนิคการถดถอยโลจิสติกมีค่าความแม่นยำ ความคลาดเคลื่อนเฉลี่ยสัมพัทธ์ และรากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ยใกล้เคียงกัน ในขณะที่เทคนิคนาอีฟเบส์มีค่าความแม่นยำต่ำ มีความคลาดเคลื่อนเฉลี่ยสัมพัทธ์ และรากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ยสูงกว่า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 3
วิธีการดำเนินงานวิจัย

3.1 การเก็บรวบรวมข้อมูล

ทำการค้นหาและศึกษาข้อมูลที่มีค่านอกเกณฑ์จากเว็บไซต์ UCI จากการศึกษาในการหาค่านอกเกณฑ์ พบว่าข้อมูลส่วนมากมีค่านอกเกณฑ์อยู่ระหว่าง 0-10 เปอร์เซ็นต์ เราจึงแบ่งข้อมูลออกเป็น 3 ระดับคือ ค่านอกเกณฑ์ระดับต่ำ 0-3.3 เปอร์เซ็นต์ ค่านอกเกณฑ์ระดับปานกลาง 3.4-6.7 เปอร์เซ็นต์ ค่านอกเกณฑ์ระดับสูง 6.8-10 เปอร์เซ็นต์ โดยได้ข้อมูล 3 ชุด คือ

1. โรคมะเร็งเต้านมของรัฐวิสคอนซิน (Breast Cancer Wisconsin) จำนวนข้อมูลทั้งหมด 699 ค่า พบค่านอกเกณฑ์จำนวน 13 ค่า คิดเป็น 1.85 เปอร์เซ็นต์ เป็นชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่ต่ำ
(จากเว็บไซต์ https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original))

2. โรคเบาหวานของชาวพิม่า ประเทศอินเดีย (Pima Indians Diabetes) จำนวนข้อมูลทั้งหมด 768 ค่า พบค่านอกเกณฑ์จำนวน 32 ค่า คิดเป็น 4.16 เปอร์เซ็นต์ เป็นชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่ปานกลาง
(จากเว็บไซต์ https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes)

3. การชำระเงินด้วยบัตรเครดิตของลูกค้า (Default Of Credit Card Clients) จำนวนข้อมูลทั้งหมด 646 ค่า พบค่านอกเกณฑ์จำนวน 62 ค่า คิดเป็น 9.6 เปอร์เซ็นต์ เป็นชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่สูง
(จากเว็บไซต์ https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients)

3.2 ตัวแปรและนิยาม

3.2.1 โรคมะเร็งเต้านมของรัฐวิสคอนซิน (Breast Cancer Wisconsin)

ตัวแปรอิสระ

X_1 แทน ความหนาของก้อนเนื้อ (หน่วย : มิลลิเมตร)

X_2 แทน ความสม่ำเสมอของขนาดเซลล์ (หน่วย : นาโนเมตร)

X_3 แทน ความสม่ำเสมอของรูปร่างเซลล์ (หน่วย : ยูนิตเซลล์)

X_4 แทน การเกาะติดขอบของเซลล์ (หน่วย : เซลล์)

X_5 แทน ขนาดเซลล์เดียว (หน่วย : มิลลิเมตร)

X_6 แทน นิวเคลียสไม่ถูกห่อหุ้ม (หน่วย : นิวเคลียส)

X_7 แทน ของเหลวในเซลล์ (หน่วย : นาโนเมตร)

X_8 แทน นิวคลีโอไลในภาวะปกติ (หน่วย : เซลล์)

X_9 แทน การขยายตัวของเซลล์ (หน่วย : เซลล์)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตัวแปรตาม

Y แทน ความเสี่ยงการเป็นโรคมะเร็งเต้านม

(No คือไม่เป็นโรคมะเร็งเต้านม และ Yes คือเป็นโรคมะเร็งเต้านม)

3.2.2 โรคเบาหวานของชาวพิม่า ประเทศอินเดีย (Pima Indians Diabetes)

ตัวแปรอิสระ

X_1 แทน จำนวนครั้งของหญิงที่ตั้งครรภ์มาแล้ว (หน่วย : ครั้ง)

X_2 แทน ความเข้มข้นของกลูโคสในเลือด 2 ชั่วโมงก่อนการทดสอบระดับกลูโคสในช่องปาก (หน่วย : มิลลิกรัมต่อเดซิเบล)

X_3 แทน ความดันโลหิตสูง (หน่วย : มิลลิเมตรปรอท)

X_4 แทน ความหนาของชั้นไขมันใต้ผิวหนังบริเวณกล้ามที่ด้านหน้าของต้นแขน (หน่วย : มิลลิเมตร)

X_5 แทน จำนวนอินซูลิน หลังจากอดอาหารภายใน 2 ชั่วโมง (หน่วย : 1,000,000 ยูนิตต่อมิลลิลิตร)

X_6 แทน ค่าดัชนีมวลกาย (หน่วย : กิโลกรัมต่อตารางเมตร)

X_7 แทน พันธุกรรมของผู้ป่วยที่มีความเสี่ยงเป็นโรคเบาหวาน (หน่วย : ร้อยละ)

X_8 แทน อายุ (หน่วย : ปี)

ตัวแปรตาม

Y แทน ความเสี่ยงการเป็นโรคเบาหวาน

(No คือ ไม่เป็นโรคเบาหวาน และ Yes คือ เป็นโรคเบาหวาน)

3.2.3 การชำระเงินด้วยบัตรเครดิตของลูกค้า (Default of Credit Card Clients)

ตัวแปรอิสระ

X_1 แทน จำนวนเครดิตที่ได้รับ รวมทั้งสินเชื่อผู้บริโภคแต่ละรายและเครดิตครอบครัว (หน่วย : สกุลเงินไต้หวันแบบใหม่ มีหน่วยเป็น TWD)

X_2 แทน เพศ (1 คือเพศชาย 2 คือเพศหญิง)

X_3 แทน การศึกษา (1 คือสำเร็จการศึกษา 2 คือกำลังศึกษาระดับมหาวิทยาลัย 3 คือกำลังศึกษามัธยมศึกษาตอนปลาย 4 คืออื่นๆ)

X_4 แทน สถานภาพการสมรส (1 คือแต่งงาน 2 คือโสด 3 คืออื่นๆ)

X_5 แทน อายุ (หน่วย : ปี)

X_6-X_{11} แทน ประวัติการชำระเงินรายเดือนย้อนหลัง (ตั้งแต่เมษายนถึงกันยายน 2005) ดังนี้

X_6 แทน สถานการณ์ชำระเงินในเดือนกันยายน

X_7 แทน สถานการณ์ชำระเงินในเดือนสิงหาคม

X_8 แทน สถานการณ์ชำระเงินในเดือนกรกฎาคม

X_9 แทน สถานการณ์ชำระเงินในเดือนมิถุนายน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

X_{10} แทน สถานการณ์ชำระเงินในเดือนพฤษภาคม

X_{11} แทน สถานการณ์ชำระเงินในเดือนเมษายน

สถานการณ์ชำระเงินคืน เป็น

-2 แทน ชำระเงินล่วงหน้า 1 เดือน -1 แทน ชำระเงินตรงเวลา

0 แทน ยังไม่ได้ชำระเงิน 1 แทน ชำระเงินล่าช้า 1 เดือน

2 แทน ชำระเงินล่าช้า 2 เดือน 3 แทน ชำระเงินล่าช้า 3 เดือน

4 แทน ชำระเงินล่าช้า 4 เดือน 5 แทน ชำระเงินล่าช้า 5 เดือน

6 แทน ชำระเงินล่าช้า 6 เดือน 7 แทน ชำระเงินล่าช้า 7 เดือนขึ้นไป

X_{12}-X_{17} แทน ใบแจ้งจำนวนค่าใช้จ่าย (ตั้งแต่เมษายนถึงกันยายน 2005) ดังนี้

X_{12} แทน จำนวนค่าใช้จ่ายในเดือนกันยายน

X_{13} แทน จำนวนค่าใช้จ่ายในเดือนสิงหาคม

X_{14} แทน จำนวนค่าใช้จ่ายในเดือนกรกฎาคม

X_{15} แทน จำนวนค่าใช้จ่ายในเดือนมิถุนายน

X_{16} แทน จำนวนค่าใช้จ่ายในเดือนพฤษภาคม

X_{17} แทน จำนวนค่าใช้จ่ายในเดือนเมษายน

X_{18}-X_{23} แทน จำนวนเงินการชำระเงินก่อนหน้านี้ (ตั้งแต่เมษายนถึงกันยายน 2005) ดังนี้

X_{18} แทน จำนวนเงินที่จ่ายในเดือนกันยายน

X_{19} แทน จำนวนเงินที่จ่ายในเดือนสิงหาคม

X_{20} แทน จำนวนเงินที่จ่ายในเดือนกรกฎาคม

X_{21} แทน จำนวนเงินที่จ่ายในเดือนมิถุนายน

X_{22} แทน จำนวนเงินที่จ่ายในเดือนพฤษภาคม

X_{23} แทน จำนวนเงินที่จ่ายในเดือนเมษายน

ตัวแปรตาม

Y แทน การชำระเงินในเดือนถัดไป (No คือ ไม่ชำระ และ Yes คือ ชำระ)

3.3 เครื่องมือที่ใช้ในการวิจัย

โปรแกรม WEKA เวอร์ชั่น 3.8 และโปรแกรม SPSS เวอร์ชั่น 22

3.4 ขั้นตอนในการดำเนินงานวิจัย

3.4.1 การศึกษาข้อกำหนดเบื้องต้นและวิธีการหาค่านอกเกณฑ์ (Outlier) ของข้อมูล

1) ติดตั้ง SQL Server 2014 ในโปรแกรม Excel

2) ติดตั้ง Data Mining Add-in สำหรับโปรแกรม Excel เพื่อสามารถใช้เครื่องมือวิเคราะห์ Data Mining เบื้องต้นในโปรแกรม Excel ได้

3) นำชุดข้อมูลที่รวบรวมได้มาทำการวิเคราะห์ค่านอกเกณฑ์ โดยใช้เครื่องมือ Highlight Exceptions จะได้จำนวนค่านอกเกณฑ์ของข้อมูลแต่ละชุด

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.4.2 การวิเคราะห์หาค่านอกเกณฑ์ (Outlier)

วิเคราะห์หาค่านอกเกณฑ์ (Outlier) จากโปรแกรม Microsoft Excel Data Mining Client สำหรับ Excel Add-in โดยใช้เครื่องมือ Highlight Exceptions ซึ่งเป็นอัลกอริทึมการจัดกลุ่มของ Microsoft รูปแบบการจัดกลุ่มจะตรวจจับกลุ่มของแถวที่มีลักษณะคล้ายกัน

3.4.3 วิธีการแบ่งข้อมูล

3.4.3.1 ทำการแบ่งชุดข้อมูลโดยโปรแกรม SPSS เวอร์ชั่น 22 ด้วยวิธีการสุ่มตัวอย่างอย่างง่าย (Simple Random Sampling: SRS) ทำการสุ่มจำนวน 3 รอบ โดยการกำหนดตัวเลขสุ่มเทียม (RandomSeed) เป็น 10,20,30 ในอัตราส่วน 70:20:10 (พยูน พาณิชย์กุล, 2548) ส่วนที่ 1 ข้อมูลเรียนรู้ (Training data) นำไปสร้างตัวแบบ (Model) 70 เปอร์เซ็นต์ ข้อมูลส่วนที่ 2 ข้อมูลตรวจสอบ (Validation data) นำไปประเมินความผิดพลาดของตัวแบบ 20 เปอร์เซ็นต์ และข้อมูลส่วนที่ 3 ข้อมูลทดสอบ (Testing data) นำไปทดสอบตัวแบบ 10 เปอร์เซ็นต์

1) เปิดไฟล์ชุดข้อมูลนามสกุล .CSV เข้าสู่โปรแกรม SPSS เวอร์ชั่น 23 โดยคลิก File ▸ Open ▸ Data ▸ เลือกไฟล์ชุดข้อมูล .CSV

2) ทำการกำหนดค่า Seed ▸ Transforn ▸ Random Number Generators ▸ ในช่อง Active Generator ▸ เลือก Set adtive Geneertor ▸ เลือก SPSS 12 Compatible

3) ในช่อง Active Generator Initialzation ▸ เลือก Set Starting Point ▸ เลือก Fixed Value ▸ ในช่อง Value กำหนดเป็น 10, 20, 30 ▸ OK

4) เลือก Data ▸ Select Case ▸ Random Sample of case Sample ▸ Approximately ใส่จำนวนเปอร์เซ็นต์ที่ต้องการแบ่งข้อมูลชุดแรก (ในงานวิจัยนี้ใส่ 70%) ▸ Continue ▸ OK

5) เมื่อดูในหน้าต่าง Data View จะเห็นว่าในคอลัมน์ filter_$ จะมีหมายเลข 1 จำนวน 70% ตามที่กำหนดไว้ข้างต้นเพื่อนำไปสร้างตัวแบบและที่เหลือหมายเลข 0 อีก 30% เพื่อนำไปทดสอบตัวแบบ บันทึกข้อมูลที่แบ่งได้ นั่นคือ 70% และ 30%

6) นำข้อมูล 30% มาแบ่งอีกรอบเพื่อแบ่งเป็น 20%-10% เลือก Data ▸ Select Case ▸ Random Sample of case Sample ▸ Approximately ใส่จำนวนเปอร์เซ็นต์ที่ต้องการแบ่งข้อมูลชุดแรก (ในงานวิจัยนี้ใส่ 20%) ▸ Continue ▸ OK

7) เมื่อดูในหน้าต่าง Data View จะเห็นว่าในคอลัมน์ filter_$ จะมีหมายเลข 1 จำนวน 20% ตามที่กำหนดไว้ข้างต้นเพื่อตรวจสอบความถูกต้องของข้อมูลและที่เหลือหมายเลข 0 อีก 10% เพื่อนำไปทดสอบตัวแบบ บันทึกข้อมูลที่แบ่งได้ นั่นคือ 20% และ 10%

3.4.3.2 ทำการแบ่งชุดข้อมูลโดยโปรแกรม WEKA เวอร์ชั่น 3.8 ทำการสุ่มจำนวน 3 รอบ โดยการกำหนดตัวเลขสุ่มเทียม (Random Seed) เป็น 10, 20, 30 ในอัตราส่วน 70:20:10 ส่วนที่ 1 ข้อมูลเรียนรู้ (Training data) นำไปสร้างตัวแบบ (Model) 70 เปอร์เซ็นต์ ข้อมูลส่วนที่ 2 ข้อมูล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตรวจสอบ (Validation data) นำไปประเมินความผิดพลาดของตัวแบบ 20 เปอร์เซ็นต์ และข้อมูลส่วนที่ 3 ข้อมูลทดสอบ (Testing data) นำไปทดสอบตัวแบบ 10 เปอร์เซ็นต์ โดยดูวิธีตามที่มาต่อไปนี้ https://www.youtube.com/watch?v=uiDFa7iY9yo

3.4.4 การศึกษาอัลกอริทึม

3.4.4.1 วิธีนาอีฟเบส์ (Naive Bayes)

นาอีฟเบส์ใช้การคำนวณหาความน่าจะเป็นซึ่งถูกใช้ในการทำนายผล เป็นวิธีในการแก้ปัญหาแบบจำแนกที่สามารถคาดการณ์ผลลัพธ์ จะทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเพื่อใช้ในการสร้างเงื่อนไขความน่าเป็นสำหรับแต่ละความสัมพันธ์ นาอีฟเบส์เป็นวิธีจำแนกกลุ่มข้อมูลที่มีประสิทธิภาพ มีอัลกอริทึมในการใช้งานที่ไม่ซับซ้อน เหมาะกับกรณีของเซตตัวอย่างที่มีจำนวนมากและคุณลักษณะ (attribute) ของตัวอย่างไม่ขึ้นต่อกัน (วรรณสิริ ธุระชนและคณะ, 2557)

3.4.4.2 วิธีเพื่อนบ้านใกล้สุด k ตัว (k-Nearest Neighbor)

เป็นวิธีการที่ได้รับความนิยมอย่างมาก เนื่องจากเป็นวิธีการที่ง่ายและมีประสิทธิภาพ ซึ่งสามารถนำไปประยุกต์ใช้กับงานได้อย่างหลากหลาย เช่น งานทางด้านการจำแนกกลุ่ม (Classification) รวมถึงงานทางด้านการแทนที่ข้อมูลที่สูญหาย (Missing Values Imputation) ใช้อัลกอริทึม IBK (O. Troyanskaya, 2001)

3.4.4.3 วิธีต้นไม้ตัดสินใจ (Decision Tree)

ต้นไม้ที่ใช้ในการสนับสนุนการตัดสินใจ ซึ่งมีลักษณะเป็นโครงสร้างต้นไม้หัวกลับที่มีรากอยู่ด้านบนและใบอยู่ด้านล่างสุด โดยที่ภายในต้นไม้จะประกอบไปด้วยโหนด (Node) ซึ่งแต่ละโหนดนั้นจะแสดงถึงการตัดสินใจบนข้อมูลของคุณลักษณะต่างๆ กิ่งของต้นไม้แสดงถึงค่าหรือผลลัพธ์ที่ได้จากการทดสอบและใบซึ่งเป็นสิ่งที่อยู่ล่างสุดของต้นไม้ตัดสินใจจะแสดงถึงกลุ่มของข้อมูล (Class) หรือผลลัพธ์โหนดที่อยู่บนสุดจะเรียกว่า โหนดราก (Root Node) ใช้อัลกอริทึม J48 (C4.5) (รุจิรา ธรรมสมบัติ, 2554)

3.4.4.4 วิธีโครงข่ายประสาทเทียม (Artificial Neural Network)

เป็นการส่งผ่านข้อมูลระหว่างกันโดยมีการเชื่อมต่อของเซลล์ประสาท (Neuron) กันเป็นโครงข่ายร่างแหจำนวนมาก สาเหตุหลักที่โครงข่ายประสาทเทียมเป็นที่นิยมกันมากขึ้นเนื่องจากมีความยืดหยุ่นในการทำงานสูงและสามารถปรับตัวเองให้ทำงานในสภาพที่เปลี่ยนแปลง อีกทั้งไม่จำเป็นต้องทราบตัวแบบทางคณิตศาสตร์ (Mathematical model) ที่แน่นอนของกระบวนการ เพียงแต่ชุดข้อมูลที่ประกอบด้วยข้อมูลนำเข้า (Input Data) และข้อมูลเป้าหมาย (Target Data) ของกระบวนการในจำนวนมากพอที่ใช้ในการสอน (Training) โครงข่ายประสาทเทียม ใช้อัลกอริทึมเพอร์เซปตรอนแบบหลายชั้น (วาทินี นุ้ยเพียร, 2553)

3.4.4.5 วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)

เป็นอัลกอริทึมที่สามารถนำมาช่วยแก้ปัญหาการจำแนกข้อมูล ใช้ในการวิเคราะห์ข้อมูลและจำแนกข้อมูล โดยอาศัยหลักการของการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลที่

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ถูกป้อนเข้าสู่กระบวนการสอนให้ระบบเรียนรู้ โดยเน้นไปยังเส้นแบ่งแยกกลุ่มข้อมูลได้ดีที่สุด โดยใช้อัลกอริทึม SMO ชนิดโพลิโนเมียลเคอร์เนล (จิรา แก้วสุวรรณ์, 2549)

หลังจากนั้นนำข้อมูลที่แบ่งออกเป็น 3 ส่วน มาทำการวิเคราะห์โดยใช้โปรแกรม WEKA ซึ่งทำการวิเคราะห์จากวิธีการจำแนกกลุ่มทั้ง 5 วิธี ข้างต้น

3.4.5 การเปรียบเทียบประสิทธิภาพของวิธีการจำแนกกลุ่ม

นำผลการวิเคราะห์ของแต่ละวิธีทั้ง 5 วิธี มาเปรียบเทียบประสิทธิภาพ

3.4.5.1 เมตริกซ์ความสับสน (Confusion Matrix) เป็นรูปแบบตารางที่เฉพาะเจาะจงที่นำผลลัพธ์จากการทำนายมาใส่ในรูปตารางเมตริกซ์ ซึ่งจะช่วยให้ง่ายต่อการมองเห็นค่าทำนายของอัลกอริทึม

ค่าการจำแนกได้ถูกต้อง (Correctiy Classified Instances) คือค่าที่บอกว่าชุดข้อมูลมีอัตราการทำนายถูกต้องและผิดพลาดเท่าไร

3.4.5.2 ค่าความถูกต้อง (Accuracy) คือการแสดงการวัดที่ได้มีความถูกต้องในรูปอัตราส่วน

3.4.5.3 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error : MSE) เป็นมาตรวัดการประเมินค่าได้ดี เนื่องจากความคลาดเคลื่อนกำลังสองเฉลี่ยประกอบด้วยทั้งความเอนเอียงและความแปรปรวน (สายชล สินสมบูรณ์ทอง, 2558)

3.4.5.4 ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย (Mean Absolute Deviation : MAD) คือ ค่าวัดความถูกต้องของการพยากรณ์ ที่วัดจากค่าความคลาดเคลื่อนโดยไม่คำนึงถึงทิศทางของความคลาดเคลื่อน MAD มีหน่วยวัดหน่วยเดียวกับค่าสังเกต (สายชล สินสมบูรณ์ทอง, 2558)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 4
ผลการวิเคราะห์ข้อมูล

งานวิจัยครั้งนี้ผู้วิจัยใช้การจำแนกกลุ่มโดยใช้วิธีการทำเหมืองข้อมูล โดยนำชุดข้อมูลที่ค้นคว้ามาจำนวน 3 ชุด มาทำการวิเคราะห์ข้อมูล ซึ่งจะสุ่มแบ่งข้อมูลออกเป็น 3 ส่วน คือ ส่วนที่ 1 ข้อมูลเรียนรู้ (Training data) นำไปสร้างตัวแบบ (Model) 70 เปอร์เซ็นต์ ข้อมูลส่วนที่ 2 ข้อมูลตรวจสอบความถูกต้อง (Validation data) นำไปประเมินความผิดพลาดของตัวแบบ 20 เปอร์เซ็นต์ และข้อมูลส่วนที่ 3 ข้อมูลทดสอบ (Testing data) นำไปทดสอบตัวแบบ 10 เปอร์เซ็นต์ และผู้วิจัยได้นำวิธีเหล่านี้มาประยุกต์กับงานวิจัย โดยนำมาเปรียบเทียบประสิทธิภาพผลการจำแนกกลุ่มในการทำวิจัยครั้งนี้ ผู้วิจัยสนใจศึกษาการจำแนกกลุ่ม (Classification) เพื่อเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มจากค่าความถูกต้อง (Accuracy) ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error : MSE) และค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย (Mean Absolute Deviation : MAD หรือ Mean Absolute Error : MAE) ซึ่งวิธีที่ใช้ในการทดสอบครั้งนี้มี 5 วิธี ผลการวิเคราะห์ข้อมูลแบ่งเป็นหัวข้อดังนี้

4.1 การวิเคราะห์ข้อมูลเบื้องต้นของข้อมูล
4.2 ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์
4.3 ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBk
4.4 ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48
4.5 ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron
4.6 ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO
4.7 ผลการเปรียบเทียบประสิทธิภาพของวิธีการจำแนกกลุ่มโดยใช้โปรแกรม SPSS และ WEKA

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.1 การวิเคราะห์ข้อมูลเบื้องต้นของข้อมูล

ตารางที่ 4-1 การเปรียบเทียบสัดส่วนค่าความคลาดเคลื่อนของข้อมูลโรคมะเร็งเต้านมของรัฐวิสคอนซิน

วิธีการจำแนกกลุ่ม	การแบ่งข้อมูล																	
	SPSS									WEKA								
	70%			20%			10%			70%			20%			10%		
	Random seed			Random seed			Random seed			Random seed			Random seed			Random seed		
	10	20	30	10	20	30	10	20	30	10	20	30	10	20	30	10	20	30
นาอีฟเบส์	0.0491	0.0389	0.0777	0.0286	0.0571	0.0286	0.0286	0.0429	0.0429	0.0470	0.0430	0.0491	0.0286	0.0357	0.0357	0.0286	0.0571	0.0143
เพื่อนบ้านใกล้สุด k ตัว	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ต้นไม้ตัดสินใจ	0.0327	0.0225	0.0327	0.0286	0.0214	0	0.0286	0.0143	0.0143	0.0286	0.0245	0.0204	0.0143	0.0214	0.0357	0.0143	0.0571	0
โครงข่ายประสาทเทียม	0.0266	0.0204	0.0266	0.0143	0.0214	0.0286	0.0143	0	0	0.0288	0.0266	0.0225	0	0.0071	0.1430	0.0143	0.0286	0
ซัพพอร์ตเวกเตอร์แมชชีน	0.0409	0.0327	0.0348	0.0143	0.0500	0.0286	0.0286	0.0143	0.0143	0.6389	0.0389	0.0348	0.0143	0.0214	0.0286	0.0286	0.0571	0

มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ตารางที่ 4-2 การเปรียบเทียบสัดส่วนค่าความคลาดเคลื่อนของข้อมูลโรคเบาหวานของชาวพิม่าประเทศอินเดีย

วิธีการจำแนกกลุ่ม	การแบ่งข้อมูล																	
	SPSS									WEKA								
	70%			20%			10%			70%			20%			10%		
	Random seed			Random seed			Random seed			Random seed			Random seed			Random seed		
	10	20	30	10	20	30	10	20	30	10	20	30	10	20	30	10	20	30
นาอีฟเบส์	0.2272	0.2383	0.2402	0.2143	0.1498	0.2078	0.2467	0.2727	0.2597	0.2346	0.2291	0.2458	0.2468	0.2273	0.1886	0.2078	0.2338	0.1818
เพื่อนบ้านใกล้สุด k ตัว	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ต้นไม้ตัดสินใจ	0.1304	0.1769	0.1546	0.1039	0.1429	0.1429	0.1299	0.1558	0.0650	0.1117	0.2160	0.1695	0.0844	0.0844	0.1429	0.1818	0.1169	0.1299
โครงข่ายประสาทเทียม	0.2178	0.1974	0.2216	0.1429	0.1623	0.1299	0.1299	0.1558	0.1429	0.2272	0.2142	0.1974	0.1623	0.1883	0.1364	0.1169	0.1039	0.0779
ซัพพอร์ตเวกเตอร์แมชชีน	0.2253	0.2253	0.2253	0.1948	0.2078	0.2338	0.2857	0.2467	0.2727	0.2346	0.2253	0.2309	0.2208	0.2338	0.1753	0.2208	0.2078	0.1688

สถาบันวิทยบริการ จุฬาลงกรณ์มหาวิทยาลัย ในโลกโดยพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ตารางที่ 4-3 การเปรียบเทียบสัดส่วนค่าความคลาดเคลื่อนของข้อมูลการชำระเงินด้วยบัตรเครดิตของลูกค้า

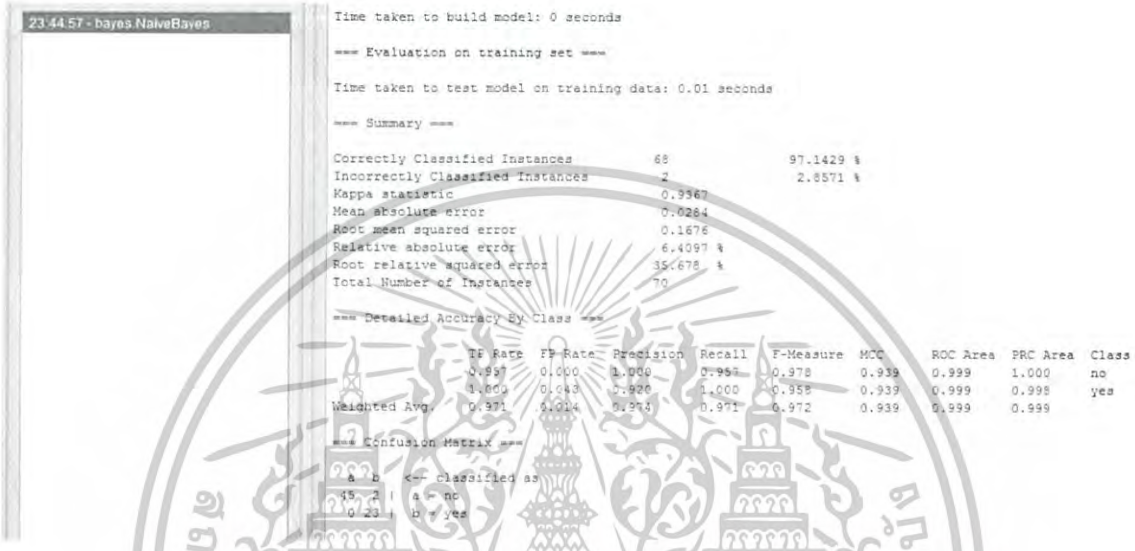
วิธีการจำแนกกลุ่ม	การแบ่งข้อมูล																	
	SPSS									WEKA								
	70%			20%			10%			70%			20%			10%		
	Random seed			Random seed			Random seed			Random seed			Random seed			Random seed		
	10	20	30	10	20	30	10	20	30	10	20	30	10	20	30	10	20	30
นาอีฟเบส์	0.4235	0.4346	0.4767	0.4264	0.4651	0.4264	0.4615	0.4462	0.4000	0.5211	0.5211	0.4279	0.4264	0.4419	0.5039	0.3231	0.2769	0.3846
เพื่อนบ้านใกล้สุด k ตัว	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ต้นไม้ตัดสินใจ	0.0710	0.0022	0.0022	0.1163	0.1318	0.0698	0.0308	0.0308	0.1078	0.0776	0.1130	0.0798	0.0233	0.0388	0.0930	0.0462	0.1077	0.0154
โครงข่ายประสาทเทียม	0.1375	0.1508	0.1375	0.1473	0.1550	0.1318	0.0769	0.0308	0.0461	0.1508	0.1486	0.1486	0.0465	0.0930	0.0620	0.0615	0.1230	0.0308
ซัพพอร์ตเวกเตอร์แมชชีน	0.1907	0.1818	0.1885	0.2481	0.2636	0.2248	0.1846	0.1385	0.1385	0.1840	0.1818	0.1840	0.1550	0.1860	0.1628	0.1538	0.1846	0.2000

จากการเปรียบเทียบสัดส่วนค่าความคลาดเคลื่อนของข้อมูลทั้ง 3 ชุด พบว่า เมื่อเปรียบเทียบการแบ่งข้อมูลของโปรแกรม SPSS และ WEKA ให้สัดส่วนค่าความคลาดเคลื่อนของข้อมูลใกล้เคียงกัน

4.2 ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์

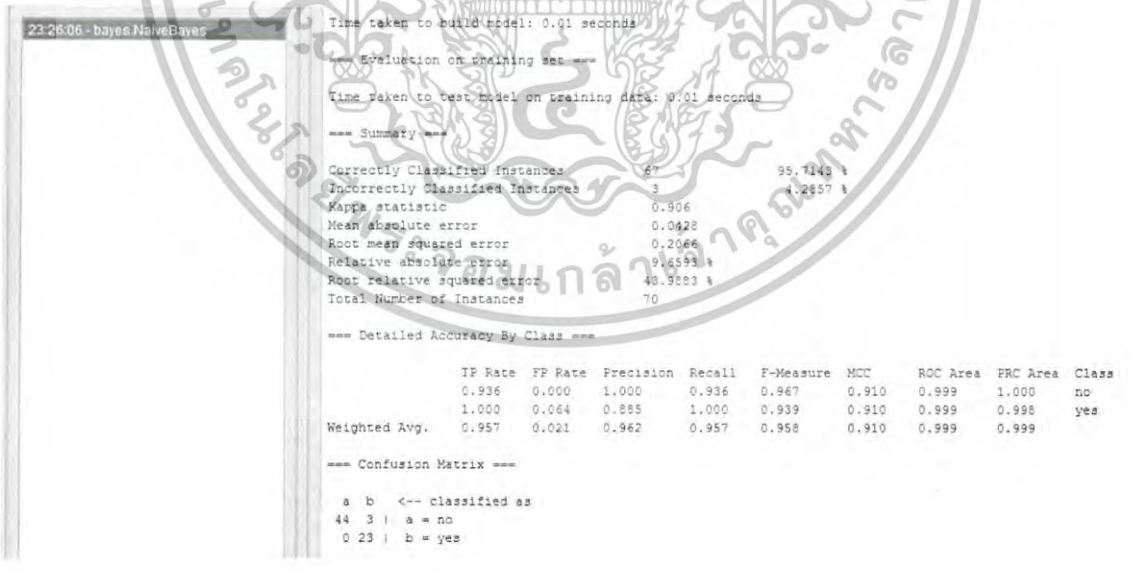
4.2.1 ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ โดยโปรแกรม SPSS

4.2.1.1 โรคมะเร็งเต้านมของรัฐวิสคอนซิน (Breast Cancer Wisconsin) จำนวนข้อมูลทดสอบ 699 ค่า ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่ต่ำ เมื่อกำหนดค่า Random Seed 10 20 และ 30 ตามลำดับ



```
23:44:57 - bayes.NaiveBayes

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances          68               97.1429 %
Incorrectly Classified Instances         2                2.8571 %
Kappa statistic                          0.9367
Mean absolute error                      0.0284
Root mean squared error                  0.1676
Relative absolute error                  6.4097 %
Root relative squared error             35.678  %
Total Number of Instances               70

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0.957     0.000     1.000     0.957     0.978     0.939     0.999     1.000     no
                 1.000     0.043     0.920     1.000     0.958     0.939     0.999     0.998     yes
Weighted Avg.    0.971     0.014     0.973     0.971     0.972     0.939     0.999     0.999

=== Confusion Matrix ===

  a  b   <-- classified as
 45  2 |  a = no
  0 23 |  b = yes
```

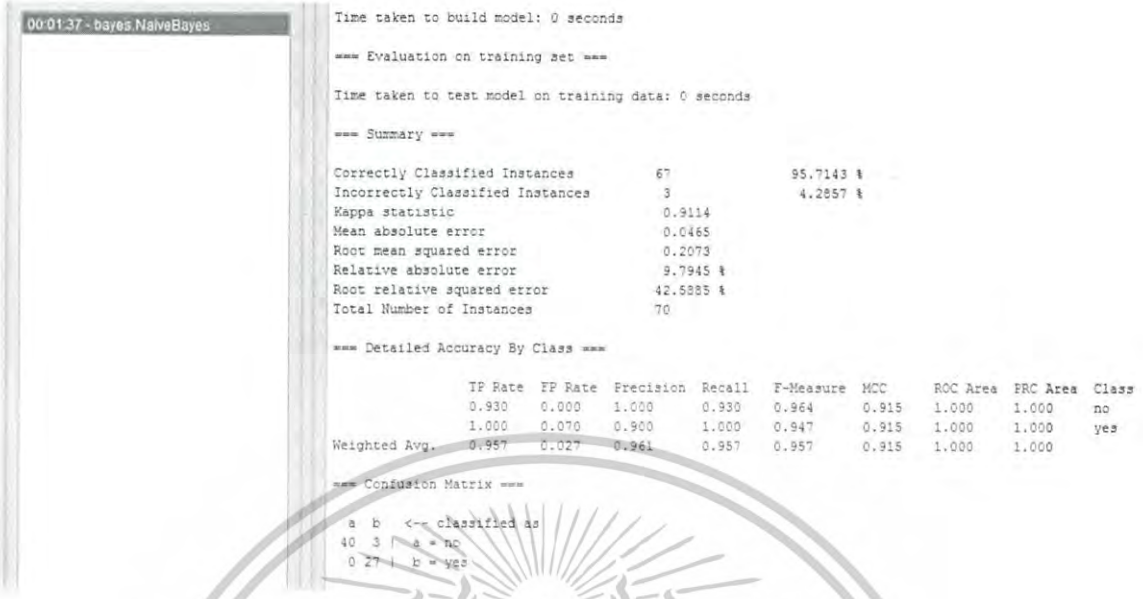
รูปที่ 4-1 ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 10



```
23:26:06 - bayes.NaiveBayes

Time taken to build model: 0.01 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances          67               95.7143 %
Incorrectly Classified Instances         3                4.2857 %
Kappa statistic                          0.906
Mean absolute error                      0.0428
Root mean squared error                  0.2066
Relative absolute error                  9.6593 %
Root relative squared error             43.9883 %
Total Number of Instances               70

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0.936     0.000     1.000     0.936     0.967     0.910     0.999     1.000     no
                 1.000     0.064     0.885     1.000     0.939     0.910     0.999     0.998     yes
Weighted Avg.    0.957     0.021     0.962     0.957     0.958     0.910     0.999     0.999

=== Confusion Matrix ===

  a  b   <-- classified as
 44  3 |  a = no
  0 23 |  b = yes
```

รูปที่ 4-2 ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 20

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



00:01:37 - bayes.NaiveBayes

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances	67	95.7143 %
Incorrectly Classified Instances	3	4.2857 %
Kappa statistic	0.9114	
Mean absolute error	0.0465	
Root mean squared error	0.2073	
Relative absolute error	9.7945 %	
Root relative squared error	42.5885 %	
Total Number of Instances	70	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.930	0.000	1.000	0.930	0.964	0.915	1.000	1.000	no
	1.000	0.070	0.900	1.000	0.947	0.915	1.000	1.000	yes
Weighted Avg.	0.957	0.027	0.961	0.957	0.957	0.915	1.000	1.000	

=== Confusion Matrix ===

a	b	<-- classified as
40	3	a = no
0	27	b = yes

รูปที่ 4-3 ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 30

จากรูป 4-1 ถึง 4-3 ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ โดยโปรแกรม SPSS โรคมะเร็งเต้านมของรัฐวิสคอนซิน โดยทำการสุ่ม 3 ครั้ง Random Seed 10 20 และ 30 ตามลำดับ จากผลลัพธ์ที่ได้พบว่า

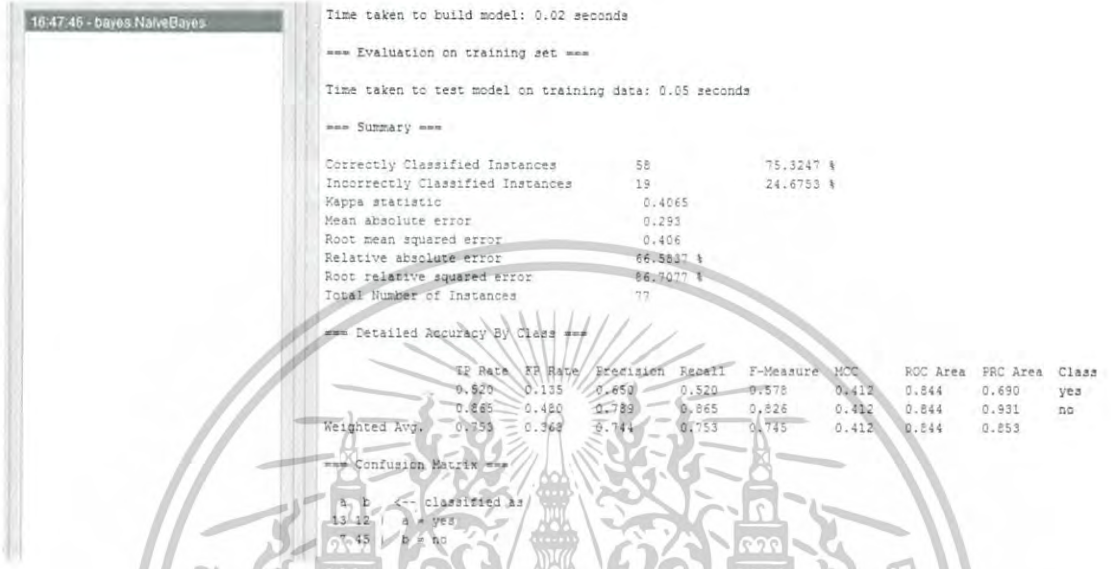
สุ่มครั้งที่ 1 Random Seed = 10 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 70 คน มีค่าทำนายความถูกต้องจำนวน 68 คน คิดเป็น 97.1429% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.1676)^2$ = 0.0281 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0284

สุ่มครั้งที่ 2 Random Seed = 20 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 70 คน มีค่าทำนายความถูกต้องจำนวน 67 คน คิดเป็น 95.7143% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.2066)^2$ = 0.0427 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0428

สุ่มครั้งที่ 3 Random Seed = 30 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 70 คน มีค่าทำนายความถูกต้องจำนวน 67 คน คิดเป็น 95.7143% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.2073)^2$ = 0.0430 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0465

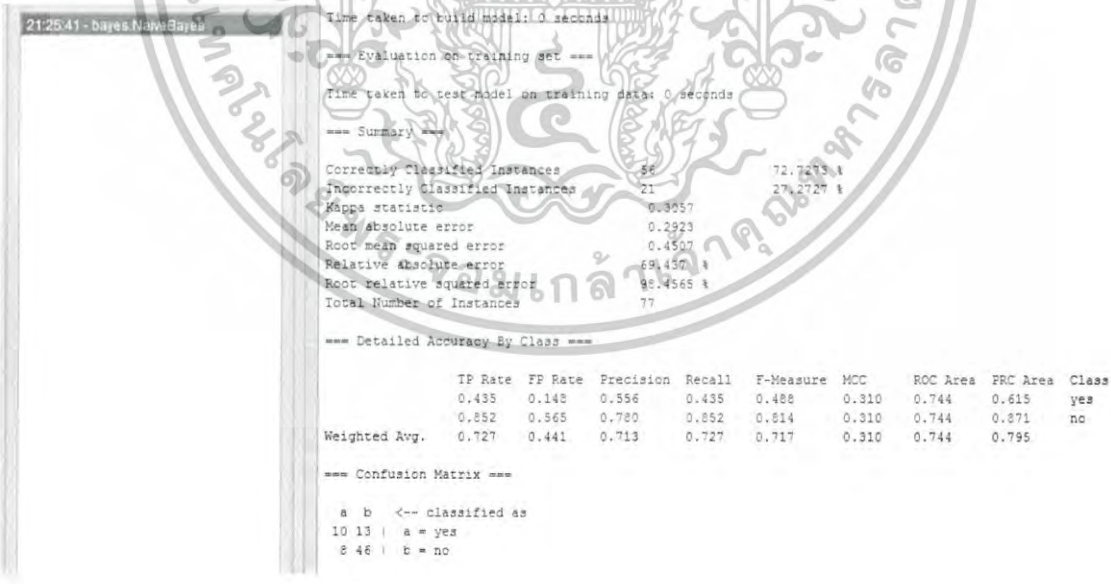
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.2.1.2 โรคเบาหวานของชาวพิม่า ประเทศอินเดีย (Pima Indians Diabetes) จำนวนข้อมูลทดสอบ 768 ค่า ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่ปานกลาง เมื่อกำหนดค่า Random Seed 10 20 และ 30 ตามลำดับ



```
16:47:46 - bayes.NaiveBayes

Time taken to build model: 0.02 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.05 seconds

=== Summary ===

Correctly Classified Instances          58               75.3247 %
Incorrectly Classified Instances        19               24.6753 %
Kappa statistic                          0.4065
Mean absolute error                      0.293
Root mean squared error                  0.406
Relative absolute error                 66.5837 %
Root relative squared error             86.7077 %
Total Number of Instances               77

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0.520     0.135     0.650     0.520     0.578     0.412     0.844     0.690     yes
                 0.865     0.480     0.789     0.865     0.826     0.412     0.844     0.931     no
Weighted Avg.    0.753     0.368     0.744     0.753     0.745     0.412     0.844     0.853

=== Confusion Matrix ===

  a  b   <-- classified as
 13 12 |  a = yes
  7 45 |  b = no
```

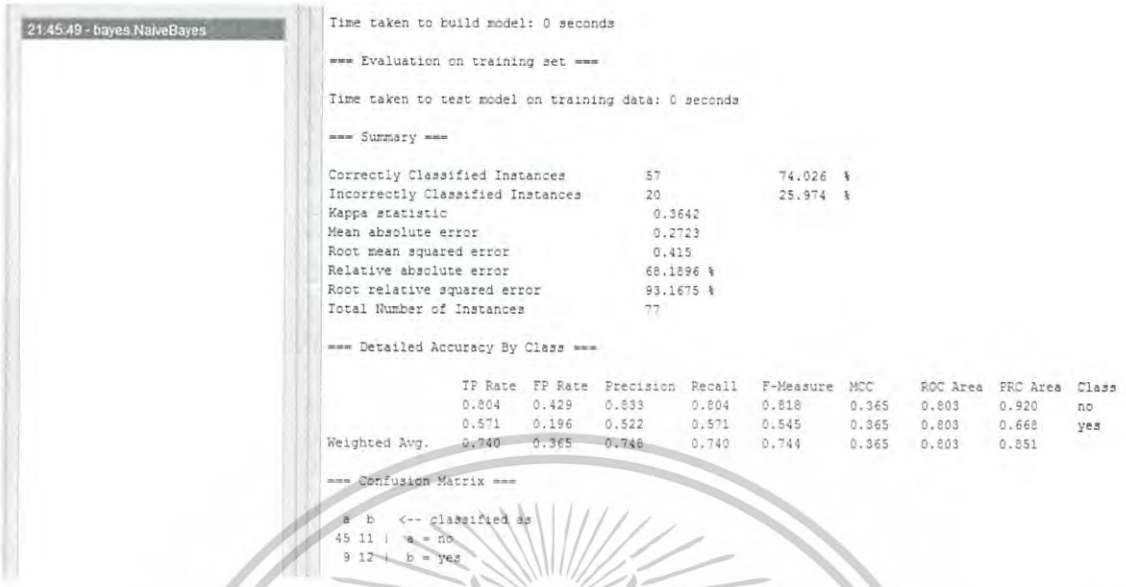
รูปที่ 4-4 ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 10



```
21:25:41 - bayes.NaiveBayes

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          56               72.7273 %
Incorrectly Classified Instances        21               27.2727 %
Kappa statistic                          0.3057
Mean absolute error                      0.2923
Root mean squared error                  0.4507
Relative absolute error                 69.4357 %
Root relative squared error             98.4565 %
Total Number of Instances               77

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0.435     0.148     0.556     0.435     0.488     0.310     0.744     0.615     yes
                 0.852     0.565     0.780     0.852     0.814     0.310     0.744     0.871     no
Weighted Avg.    0.727     0.441     0.713     0.727     0.717     0.310     0.744     0.795

=== Confusion Matrix ===

  a  b   <-- classified as
 10 13 |  a = yes
  8 46 |  b = no
```

รูปที่ 4-5 ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 20

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



21:45:49 - bayes.NaiveBayes

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances	57	74.026 %
Incorrectly Classified Instances	20	25.974 %
Kappa statistic	0.3642	
Mean absolute error	0.2723	
Root mean squared error	0.415	
Relative absolute error	68.1896 %	
Root relative squared error	93.1675 %	
Total Number of Instances	77	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.804	0.429	0.833	0.804	0.818	0.365	0.803	0.920	no
	0.571	0.196	0.522	0.571	0.545	0.365	0.803	0.668	yes
Weighted Avg.	0.740	0.365	0.748	0.740	0.744	0.365	0.803	0.851	

=== Confusion Matrix ===

a	b	<-- classified as
45	11	a = no
9	12	b = yes

รูปที่ 4-6 ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 30

จากรูป 4-4 ถึง 4-6 ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ โดยโปรแกรม SPSS โรคเบาหวานของชาวพิม่า ประเทศอินเดีย โดยทำการสุ่ม 3 ครั้ง Random Seed 10 20 และ 30 ตามลำดับ จากผลลัพธ์ที่ได้พบว่า

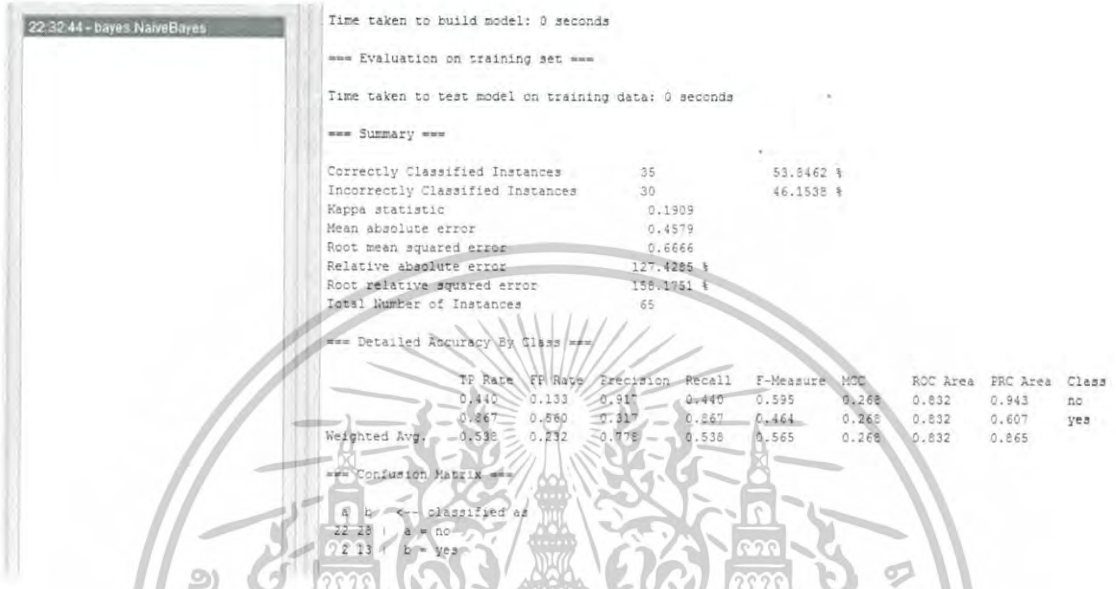
สุ่มครั้งที่ 1 Random Seed = 10 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 77 คน มีค่าทำนายความถูกต้องจำนวน 58 คน คิดเป็น 75.3247% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.406)^2$ = 0.1648 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.293

สุ่มครั้งที่ 2 Random Seed = 20 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 77 คน มีค่าทำนายความถูกต้องจำนวน 56 คน คิดเป็น 72.7273% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.4507)^2$ = 0.2031 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.2923

สุ่มครั้งที่ 3 Random Seed = 30 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 77 คน มีค่าทำนายความถูกต้องจำนวน 57 คน คิดเป็น 74.026% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.415)^2$ = 0.1722 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.2723

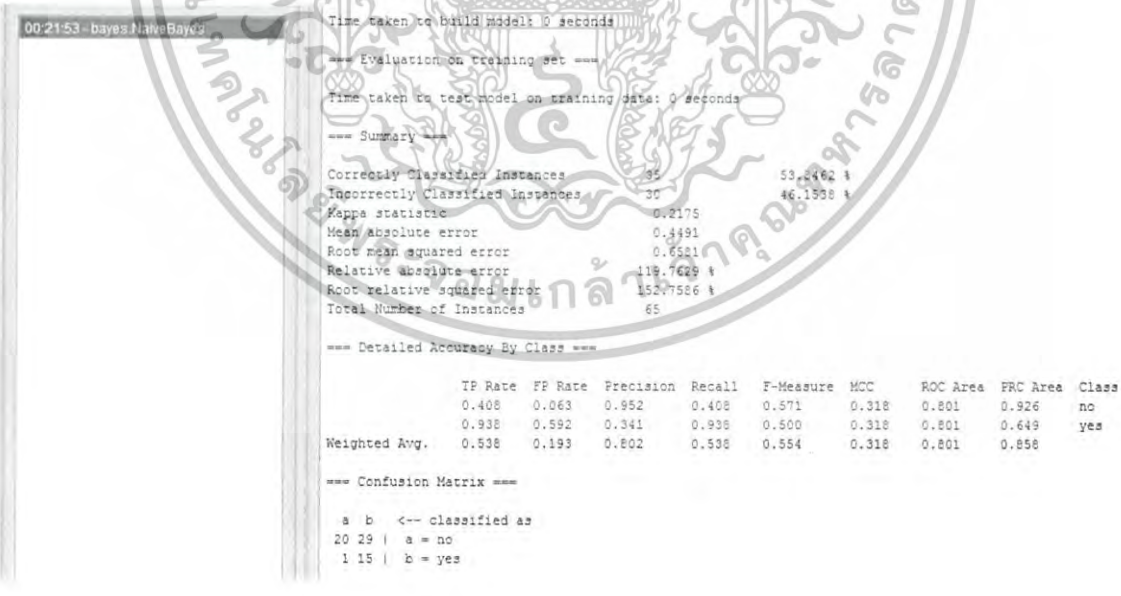
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.2.1.3 การชำระเงินด้วยบัตรเครดิตของลูกค้า (Default of Credit Card Clients) จำนวนข้อมูลทดสอบ 646 ค่า ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่สูง เมื่อกำหนดค่า Random Seed 10 20 และ 30 ตามลำดับ



```
22:32:44 - bayes.NaiveBayes

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          35               53.8462 %
Incorrectly Classified Instances        30               46.1538 %
Kappa statistic                          0.1909
Mean absolute error                      0.4579
Root mean squared error                  0.6666
Relative absolute error                127.4285 %
Root relative squared error            158.1751 %
Total Number of Instances               65

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0.440     0.133     0.917     0.440     0.595     0.268     0.832     0.943     no
                 0.867     0.560     0.317     0.867     0.464     0.268     0.832     0.607     yes
Weighted Avg.    0.538     0.232     0.778     0.538     0.565     0.268     0.832     0.865

=== Confusion Matrix ===

  a  b   <-- classified as
 22 28 |  a = no
  2 13 |  b = yes
```

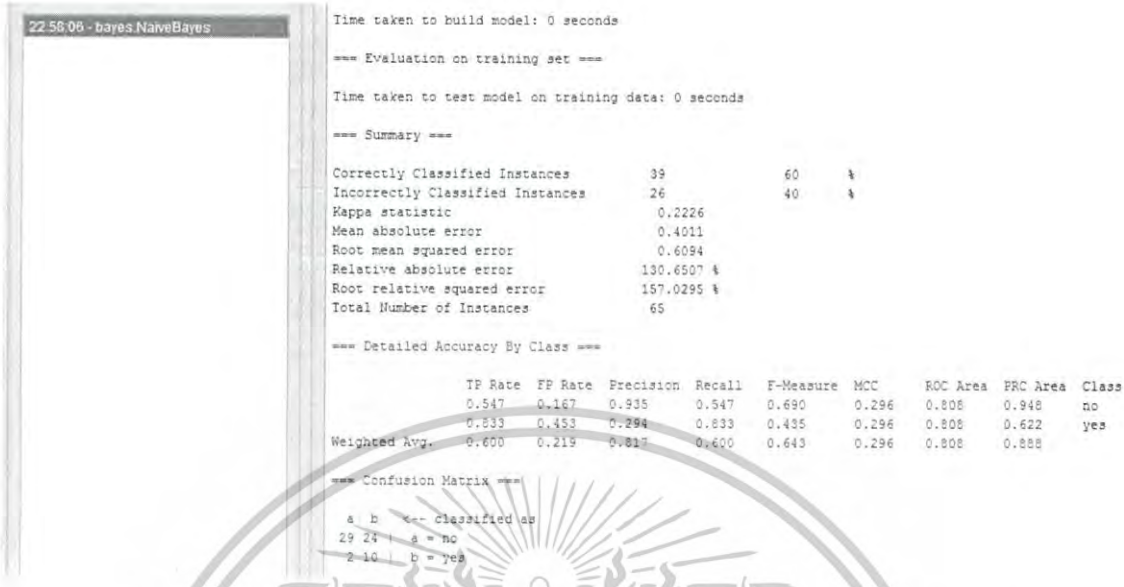
รูปที่ 4-7 ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 10



```
00:21:53 - bayes.NaiveBayes

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          35               53.8462 %
Incorrectly Classified Instances        30               46.1538 %
Kappa statistic                          0.2175
Mean absolute error                      0.4491
Root mean squared error                  0.6521
Relative absolute error                124.9629 %
Root relative squared error            152.7586 %
Total Number of Instances               65

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0.408     0.063     0.952     0.408     0.571     0.318     0.801     0.926     no
                 0.938     0.592     0.341     0.938     0.500     0.318     0.801     0.649     yes
Weighted Avg.    0.538     0.193     0.802     0.538     0.554     0.318     0.801     0.858

=== Confusion Matrix ===

  a  b   <-- classified as
 20 29 |  a = no
  1 15 |  b = yes
```

รูปที่ 4-8 ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 20

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



22:56:06 - bayes.NaiveBayes

```
Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          39               60      %
Incorrectly Classified Instances        26               40      %
Kappa statistic                          0.2226
Mean absolute error                      0.4011
Root mean squared error                  0.6094
Relative absolute error                130.6507 %
Root relative squared error            157.0295 %
Total Number of Instances               65

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0.547     0.167     0.935     0.547     0.690      0.296     0.808     0.948     no
                 0.833     0.453     0.294     0.833     0.435      0.296     0.808     0.622     yes
Weighted Avg.    0.600     0.219     0.817     0.600     0.643      0.296     0.808     0.888

=== Confusion Matrix ===

  a  b   <-- classified as
 29 24 |  a = no
  2 10 |  b = yes
```

รูปที่ 4-9 ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 30

จากรูป 4-7 ถึง 4-9 ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ โดยโปรแกรม SPSS การชำระเงินด้วยบัตรเครดิตของลูกค้า โดยทำการสุ่ม 3 ครั้ง Random Seed 10 20 และ 30 ตามลำดับ จากผลลัพธ์ที่ได้พบว่า

สุ่มครั้งที่ 1 Random Seed = 10 กลุ่มลูกค้า 10 เปอร์เซ็นต์ จำนวนทั้งหมด 65 คน มีค่าทำนายความถูกต้องจำนวน 35 คน คิดเป็น 53.8462% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.6666)^2$ = 0.4444 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.4579

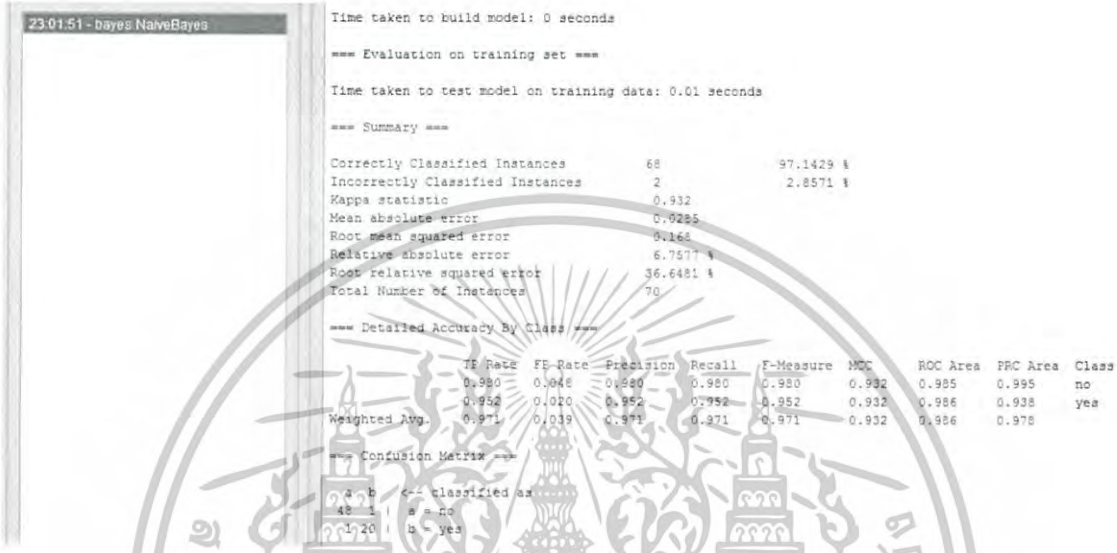
สุ่มครั้งที่ 2 Random Seed = 20 กลุ่มลูกค้า 10 เปอร์เซ็นต์ จำนวนทั้งหมด 65 คน มีค่าทำนายความถูกต้องจำนวน 35 คน คิดเป็น 53.8462% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.6581)^2$ = 0.4331 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.4491

สุ่มครั้งที่ 3 Random Seed = 30 กลุ่มลูกค้า 10 เปอร์เซ็นต์ จำนวนทั้งหมด 65 คน มีค่าทำนายความถูกต้องจำนวน 39 คน คิดเป็น 60% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.6094)^2$ = 0.3714 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.4011

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

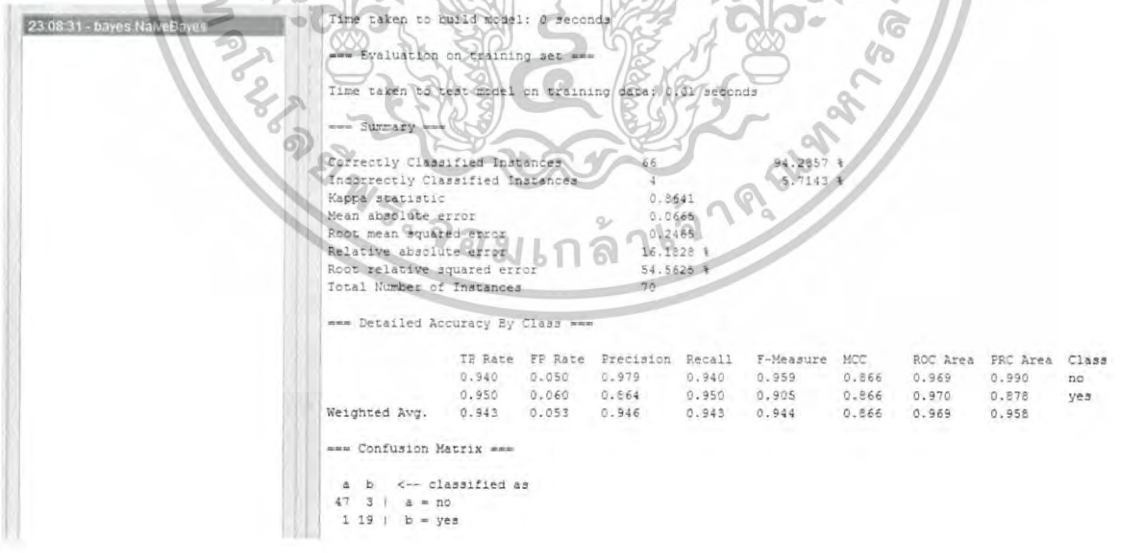
4.2.2 ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ โดยโปรแกรม WEKA

4.2.2.1 โรคมะเร็งเต้านมของรัฐวิสคอนซิน (Breast Cancer Wisconsin) จำนวนข้อมูลทดสอบ 699 ค่า ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่ต่ำ เมื่อกำหนดค่า Random Seed 10 20 และ 30 ตามลำดับ



```
23:01:51 - bayes.NaiveBayes

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances          68               97.1429 %
Incorrectly Classified Instances         2                2.8571 %
Kappa statistic                          0.932
Mean absolute error                      0.0285
Root mean squared error                  0.168
Relative absolute error                  6.7577 %
Root relative squared error             36.6481 %
Total Number of Instances               70

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.980     0.048      0.980     0.980     0.980     0.932     0.985     0.995     no
                 0.952     0.020      0.952     0.952     0.952     0.932     0.986     0.938     yes
Weighted Avg.    0.971     0.039      0.971     0.971     0.971     0.932     0.986     0.978

=== Confusion Matrix ===

  a  b   <-- classified as
 48  1 |  a = no
  1 20 |  b = yes
```

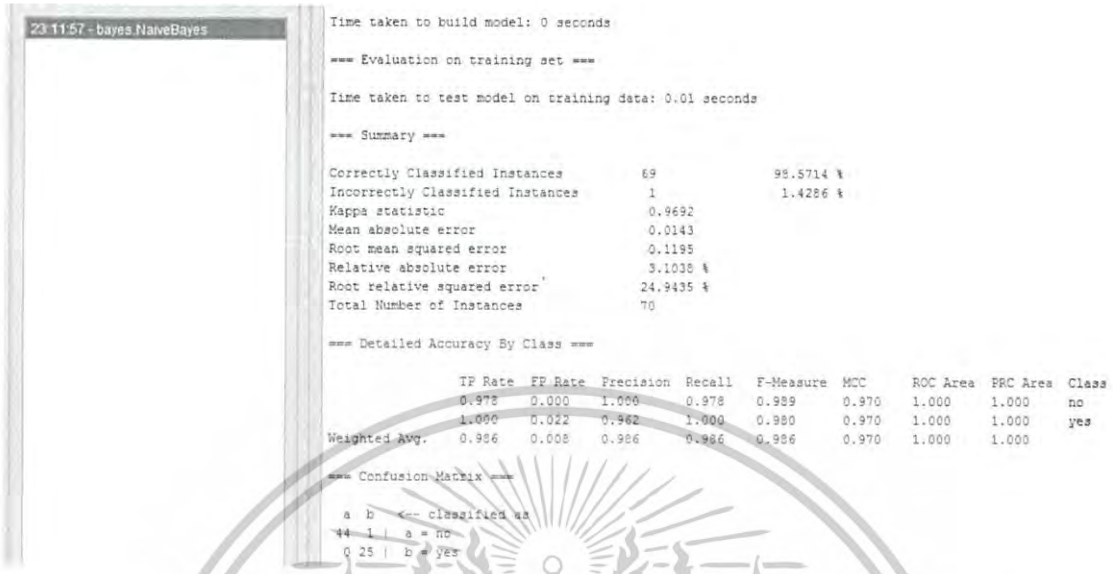
รูปที่ 4-10 ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 10



```
23:08:31 - bayes.NaiveBayes

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances          66               94.2857 %
Incorrectly Classified Instances         4                5.7143 %
Kappa statistic                          0.8641
Mean absolute error                      0.0665
Root mean squared error                  0.2465
Relative absolute error                 16.1828 %
Root relative squared error             54.5625 %
Total Number of Instances               70

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.940     0.050      0.979     0.940     0.959     0.866     0.969     0.990     no
                 0.950     0.060      0.864     0.950     0.905     0.866     0.970     0.878     yes
Weighted Avg.    0.943     0.053      0.946     0.943     0.944     0.866     0.969     0.958

=== Confusion Matrix ===

  a  b   <-- classified as
 47  3 |  a = no
  1 19 |  b = yes
```

รูปที่ 4-11 ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 20

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



23:11:57 - bayes.NaiveBayes

```
Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances          69               98.5714 %
Incorrectly Classified Instances         1                1.4286 %
Kappa statistic                          0.9692
Mean absolute error                      0.0143
Root mean squared error                  0.1195
Relative absolute error                  3.1038 %
Root relative squared error             24.9435 %
Total Number of Instances               70

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.978    0.000    1.000      0.978   0.989      0.970    1.000     1.000     no
               1.000    0.022    0.962      1.000   0.980      0.970    1.000     1.000     yes
Weighted Avg.  0.986    0.008    0.986      0.986   0.986      0.970    1.000     1.000

=== Confusion Matrix ===

  a  b   <-- classified as
 44  1 |  a = no
  0 25 |  b = yes
```

รูปที่ 4-12 ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 30

จากรูป 4-10 ถึง 4-12 ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ โดยโปรแกรม WEKA โรคมะเร็งเต้านมของรัฐวิสคอนซิน โดยทำการสุ่ม 3 ครั้ง Random Seed 10 20 และ 30 ตามลำดับ จากผลลัพธ์ที่ได้พบว่า

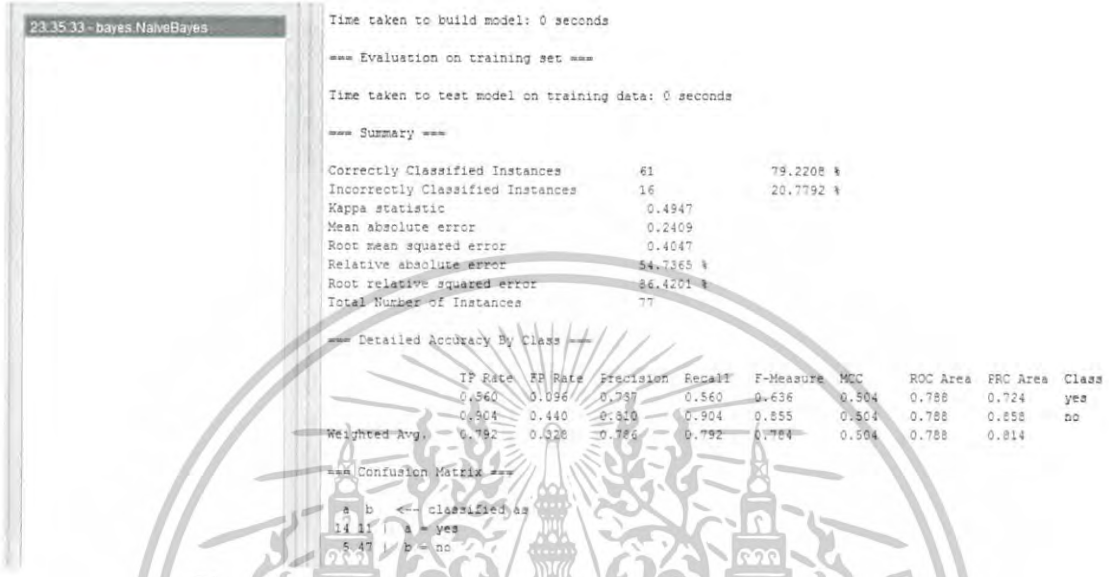
สุ่มครั้งที่ 1 Random Seed = 10 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 70 คน มีค่าทำนายความถูกต้องจำนวน 68 คน คิดเป็น 97.1429% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.168)^2$ = 0.0282 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0285

สุ่มครั้งที่ 2 Random Seed = 20 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 70 คน มีค่าทำนายความถูกต้องจำนวน 66 คน คิดเป็น 94.2857% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.2465)^2$ = 0.0608 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0665

สุ่มครั้งที่ 3 Random Seed = 30 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 70 คน มีค่าทำนายความถูกต้องจำนวน 69 คน คิดเป็น 98.5714% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.1195)^2$ = 0.0143 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0143

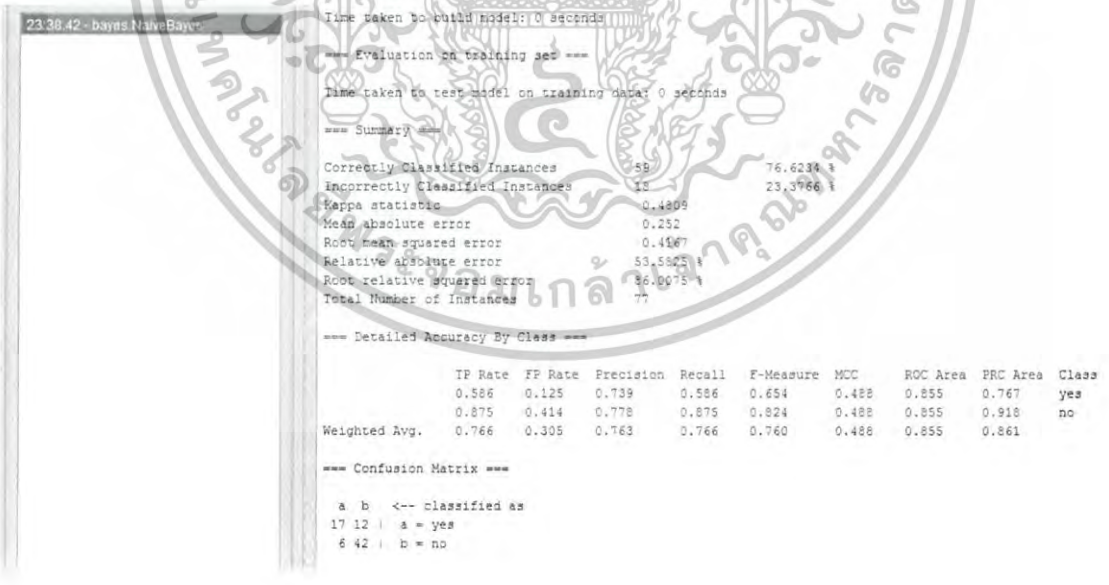
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.2.2.2 โรคเบาหวานของชาวพิม่า ประเทศอินเดีย (Pima Indians Diabetes) จำนวนข้อมูลทดสอบ 768 ค่า ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่ปานกลาง เมื่อกำหนดค่า Random Seed 10 20 และ 30 ตามลำดับ



```
23:35:33 - bayes.NaiveBayes

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          61               79.2208 %
Incorrectly Classified Instances        16               20.7792 %
Kappa statistic                          0.4947
Mean absolute error                      0.2409
Root mean squared error                  0.4047
Relative absolute error                 54.7365 %
Root relative squared error             86.4201 %
Total Number of Instances               77

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0.560     0.096     0.737     0.560     0.636     0.504     0.788     0.724     yes
                 0.904     0.440     0.810     0.904     0.855     0.504     0.788     0.858     no
Weighted Avg.    0.792     0.328     0.786     0.792     0.784     0.504     0.788     0.814

=== Confusion Matrix ===

  a  b   <-- classified as
 14 11 |  a = yes
  5 47 |  b = no
```

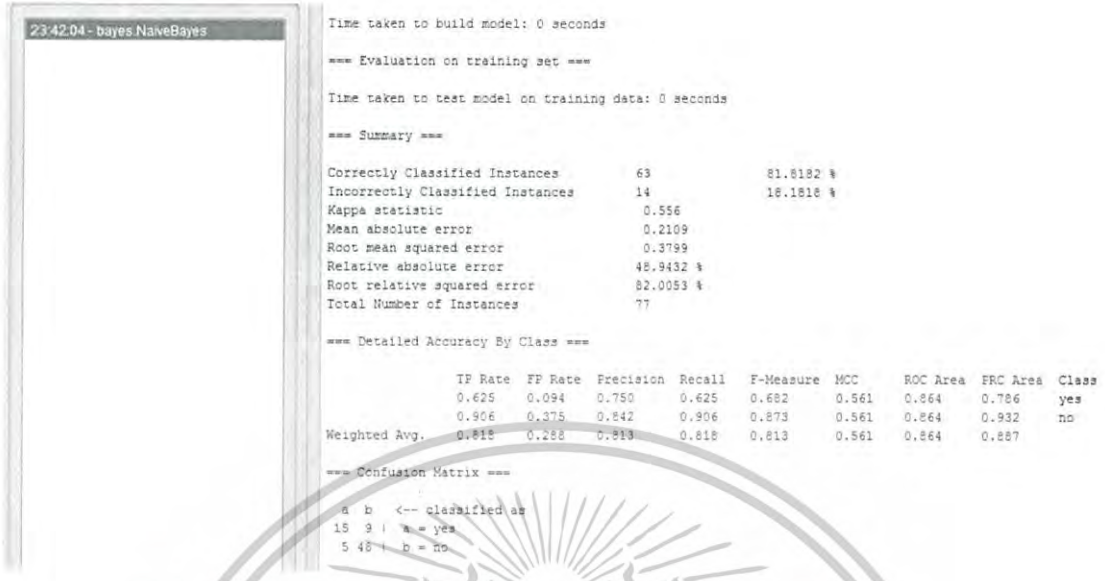
รูปที่ 4-13 ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 10



```
23:38:42 - bayes.NaiveBayes

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          59               76.6234 %
Incorrectly Classified Instances        18               23.3766 %
Kappa statistic                          0.4809
Mean absolute error                      0.252
Root mean squared error                  0.4167
Relative absolute error                 53.5825 %
Root relative squared error             86.0075 %
Total Number of Instances               77

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0.586     0.125     0.739     0.586     0.654     0.488     0.855     0.767     yes
                 0.875     0.414     0.778     0.875     0.824     0.488     0.855     0.918     no
Weighted Avg.    0.766     0.305     0.763     0.766     0.760     0.488     0.855     0.861

=== Confusion Matrix ===

  a  b   <-- classified as
 17 12 |  a = yes
  6 42 |  b = no
```

รูปที่ 4-14 ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 20

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4-15 ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 30

จากรูป 4-13 ถึง 4-15 ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ โดยโปรแกรม WEKA โรคเบาหวานของชาวพิม่า ประเทศอินเดีย โดยทำการสุ่ม 3 ครั้ง Random Seed 10 20 และ 30 ตามลำดับ จากผลลัพธ์ที่ได้พบว่า

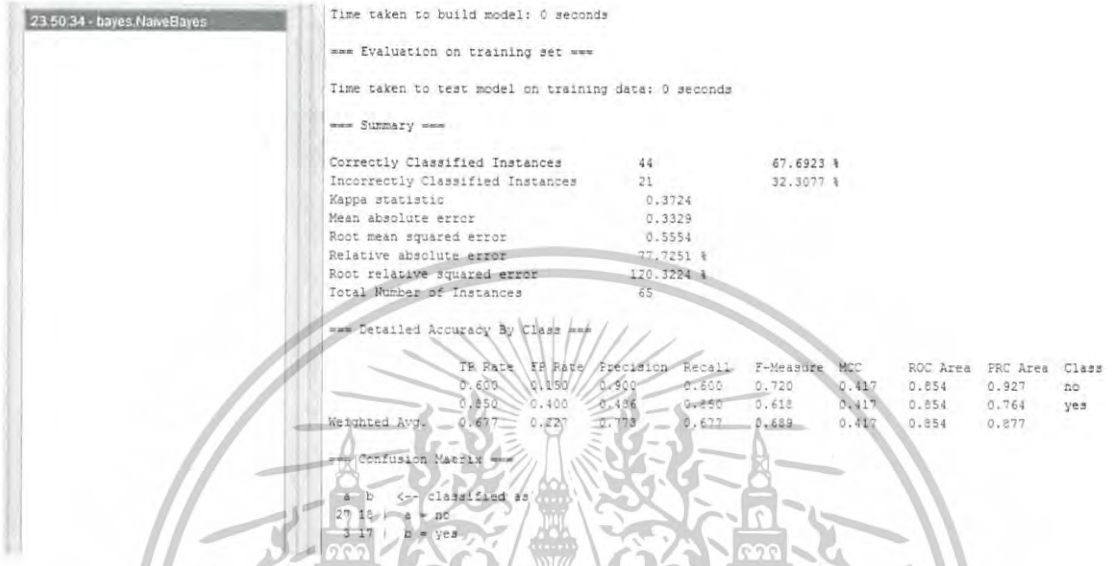
สุ่มครั้งที่ 1 Random Seed = 10 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 77 คน มีค่าทำนายความถูกต้องจำนวน 61 คน คิดเป็น 79.2208% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.4047)^2$ = 0.1638 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.2409

สุ่มครั้งที่ 2 Random Seed = 20 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 77 คน มีค่าทำนายความถูกต้องจำนวน 59 คน คิดเป็น 76.6234% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.4167)^2$ = 0.1736 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.252

สุ่มครั้งที่ 3 Random Seed = 30 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 77 คน มีค่าทำนายความถูกต้องจำนวน 63 คน คิดเป็น 81.8182% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.3799)^2$ = 0.1443 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.2109

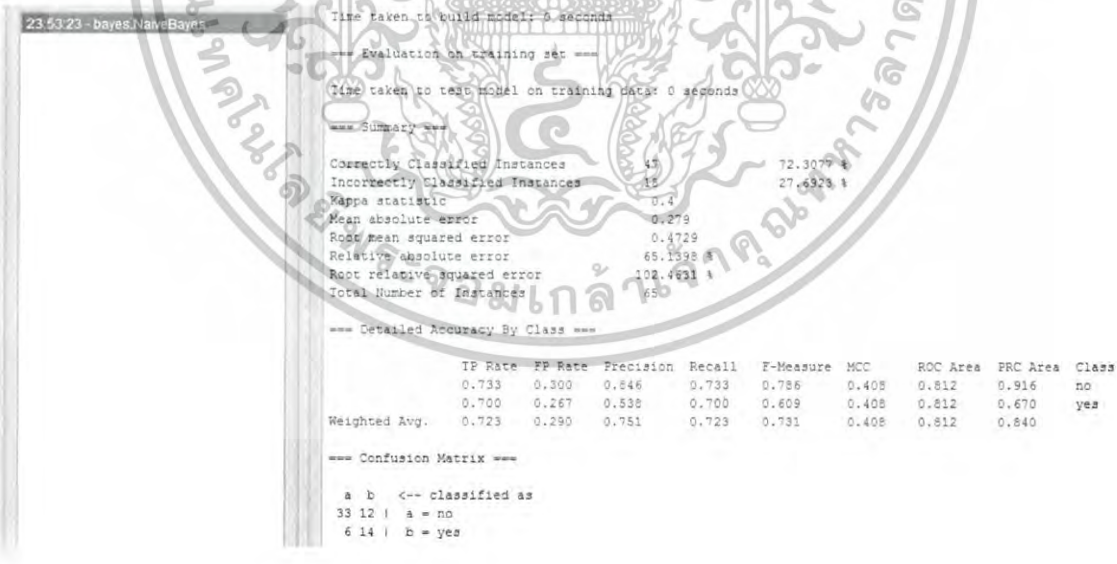
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.2.2.3 การชำระเงินด้วยบัตรเครดิตของลูกค้า (Default of Credit Card Clients) จำนวนข้อมูลทดสอบ 646 ค่า ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่สูง เมื่อกำหนดค่า Random Seed 10 20 และ 30 ตามลำดับ



```
23:50:34 - bayes.NaiveBayes

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          44               67.6923 %
Incorrectly Classified Instances        21               32.3077 %
Kappa statistic                          0.3724
Mean absolute error                      0.3329
Root mean squared error                  0.5554
Relative absolute error                 77.7251 %
Root relative squared error            120.3224 %
Total Number of Instances               65

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0.600     0.150     0.900     0.600     0.720     0.417     0.854     0.927     no
                 0.850     0.400     0.486     0.850     0.618     0.417     0.854     0.764     yes
Weighted Avg.    0.677     0.227     0.778     0.677     0.689     0.417     0.854     0.877

=== Confusion Matrix ===

  a  b   <-- classified as
 27 18 |  a = no
  3 17 |  b = yes
```

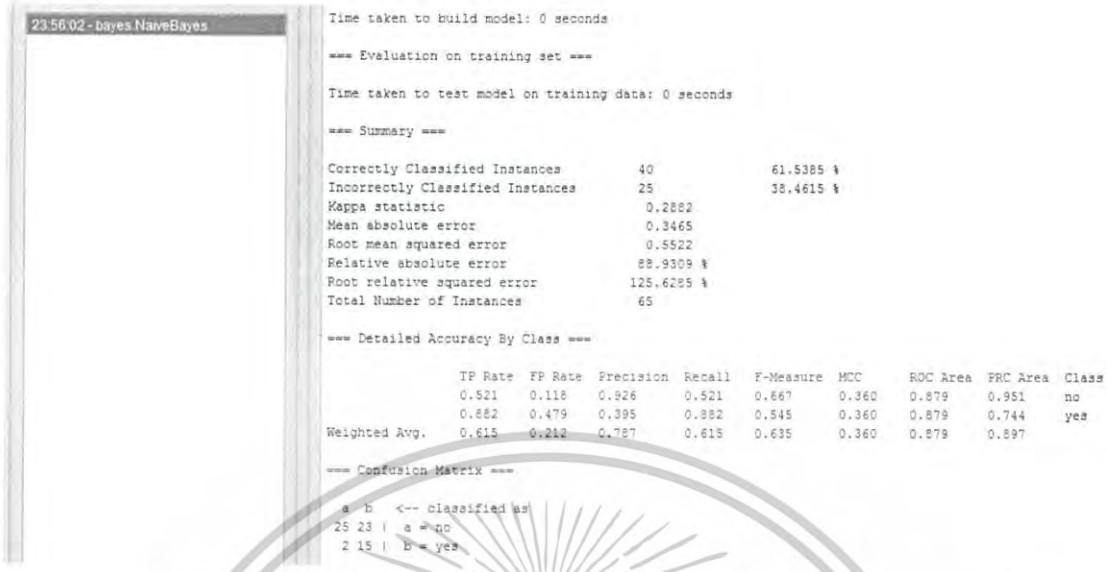
รูปที่ 4-16 ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 10



```
23:53:23 - bayes.NaiveBayes

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          47               72.3077 %
Incorrectly Classified Instances        18               27.6923 %
Kappa statistic                          0.4
Mean absolute error                      0.279
Root mean squared error                  0.4729
Relative absolute error                 65.1398 %
Root relative squared error            102.4631 %
Total Number of Instances               65

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0.733     0.300     0.846     0.733     0.786     0.408     0.812     0.916     no
                 0.700     0.267     0.538     0.700     0.609     0.408     0.812     0.670     yes
Weighted Avg.    0.723     0.290     0.751     0.723     0.731     0.408     0.812     0.840

=== Confusion Matrix ===

  a  b   <-- classified as
 33 12 |  a = no
  6 14 |  b = yes
```

รูปที่ 4-17 ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 20

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



```
23:56:02 - bayes.NaiveBayes

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          40               61.5385 %
Incorrectly Classified Instances        25               38.4615 %
Kappa statistic                          0.2882
Mean absolute error                      0.3465
Root mean squared error                  0.5522
Relative absolute error                 88.9309 %
Root relative squared error            125.6285 %
Total Number of Instances               65

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.521    0.118    0.926      0.521   0.667      0.360    0.879     0.951     no
                 0.882    0.479    0.395      0.882   0.545      0.360    0.879     0.744     yes
Weighted Avg.    0.615    0.212    0.787      0.615   0.635      0.360    0.879     0.897

=== Confusion Matrix ===

  a  b   <-- classified as
 25 23 |  a = no
  2 15 |  b = yes
```

รูปที่ 4-18 ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 30

จากรูป 4-16 ถึง 4-18 ผลการวิเคราะห์ข้อมูลด้วยวิธีนาอีฟเบส์ โดยโปรแกรม WEKA การชำระเงินด้วยบัตรเครดิตของลูกค้า โดยทำการสุ่ม 3 ครั้ง Random Seed 10 20 และ 30 ตามลำดับ จากผลลัพธ์ที่ได้พบว่า

สุ่มครั้งที่ 1 Random Seed = 10 กลุ่มลูกค้า 10 เปอร์เซ็นต์ จำนวนทั้งหมด 65 คน มีค่าทำนายความถูกต้องจำนวน 44 คน คิดเป็น 67.6923% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.5554)^2$ = 0.3085 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.3329

สุ่มครั้งที่ 2 Random Seed = 20 กลุ่มลูกค้า 10 เปอร์เซ็นต์ จำนวนทั้งหมด 65 คน มีค่าทำนายความถูกต้องจำนวน 47 คน คิดเป็น 72.3077% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.4729)^2$ = 0.2236 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.2790

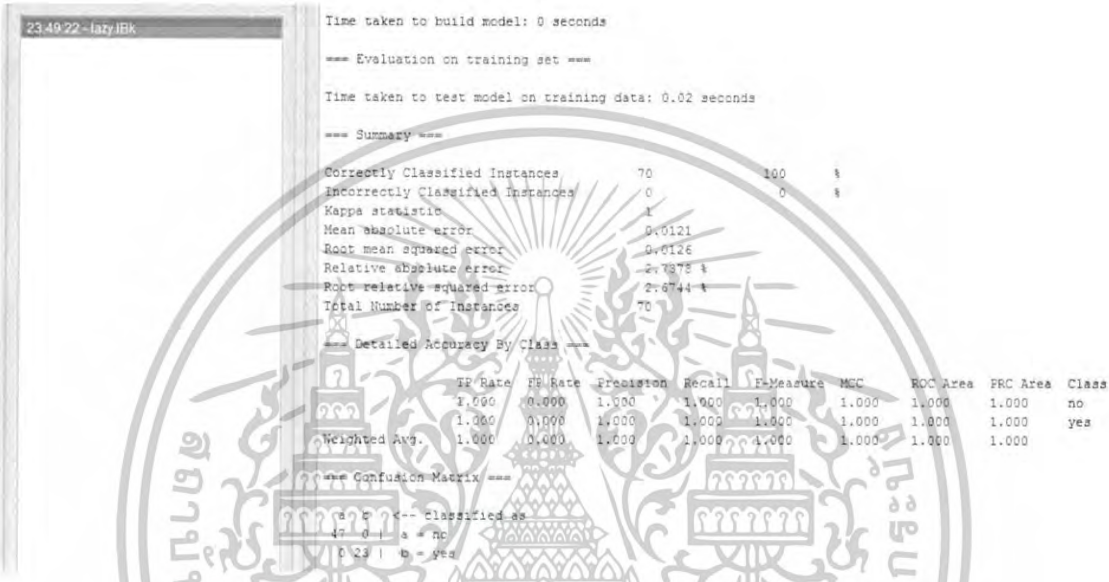
สุ่มครั้งที่ 3 Random Seed = 30 กลุ่มลูกค้า 10 เปอร์เซ็นต์ จำนวนทั้งหมด 65 คน มีค่าทำนายความถูกต้องจำนวน 40 คน คิดเป็น 61.5385% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.5522)^2$ = 0.3049 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.3465

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.3 ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBk

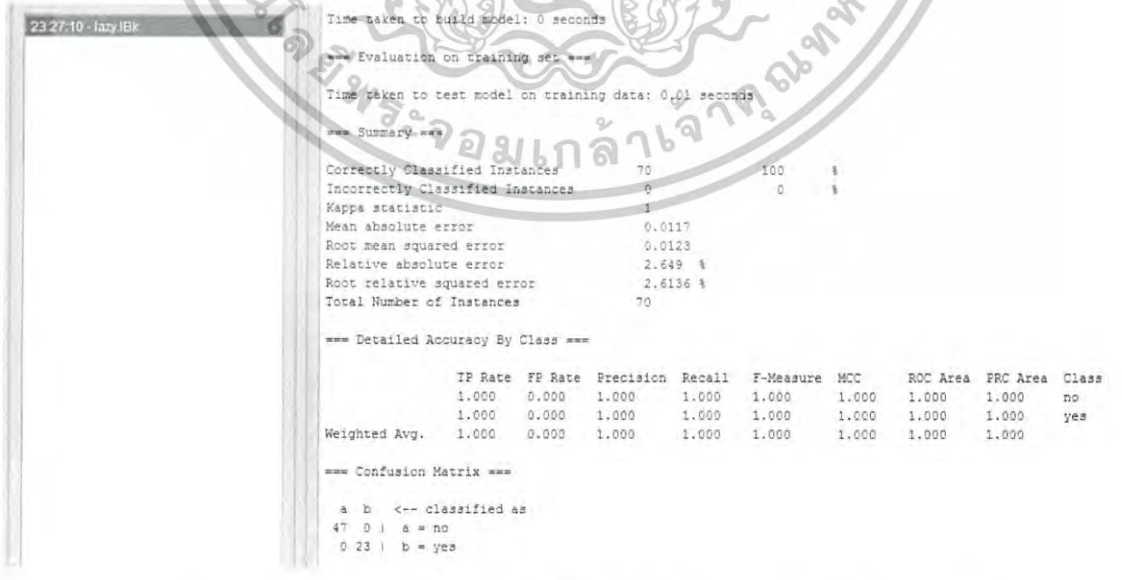
4.3.1 ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBk โดยโปรแกรม SPSS

4.3.1.1 โรคมะเร็งเต้านมของรัฐวิสคอนซิน (Breast Cancer Wisconsin) จำนวนข้อมูลทดสอบ 699 ค่า ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่ต่ำ เมื่อกำหนดค่า Random Seed 10 20 และ 30 ตามลำดับ



```
23:49:22 - lazy.IBk

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.02 seconds

=== Summary ===

Correctly Classified Instances          70               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0121
Root mean squared error                  0.0126
Relative absolute error                  2.7378 %
Root relative squared error              2.6744 %
Total Number of Instances               70

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000      1.000     1.000     1.000     1.000     1.000     1.000     no
                 1.000    0.000      1.000     1.000     1.000     1.000     1.000     1.000     yes
Weighted Avg.    1.000    0.000      1.000     1.000     1.000     1.000     1.000     1.000

=== Confusion Matrix ===

  a  b   <-- classified as
 47  0 |  a = no
  0 23 |  b = yes
```

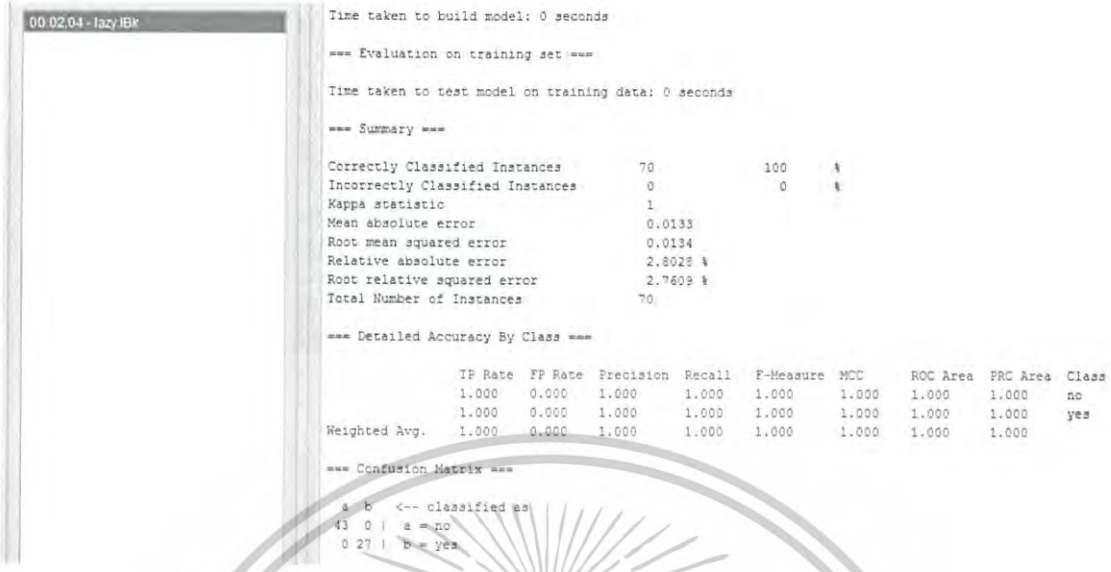
รูปที่ 4-19 ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 10



```
23:27:10 - lazy.IBk

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances          70               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0117
Root mean squared error                  0.0123
Relative absolute error                  2.649  %
Root relative squared error              2.6136 %
Total Number of Instances               70

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000      1.000     1.000     1.000     1.000     1.000     1.000     no
                 1.000    0.000      1.000     1.000     1.000     1.000     1.000     1.000     yes
Weighted Avg.    1.000    0.000      1.000     1.000     1.000     1.000     1.000     1.000

=== Confusion Matrix ===

  a  b   <-- classified as
 47  0 |  a = no
  0 23 |  b = yes
```

รูปที่ 4-20 ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 20

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



00:02:04 - lazy.IBk

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances	70	100 %
Incorrectly Classified Instances	0	0 %
Kappa statistic	1	
Mean absolute error	0.0133	
Root mean squared error	0.0134	
Relative absolute error	2.8028 %	
Root relative squared error	2.7609 %	
Total Number of Instances	70	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	no
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	yes
Weighted Avg.	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	

=== Confusion Matrix ===

a	b	<-- classified as
43	0	a = no
0	27	b = yes

รูปที่ 4-21 ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 30

จากรูป 4-19 ถึง 4-21 ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK โดยโปรแกรม SPSS โรคมะเร็งเต้านมของรัฐวิสคอนซิน โดยทำการสุ่ม 3 ครั้ง Random Seed 10 20 และ 30 ตามลำดับ จากผลลัพธ์ที่ได้พบว่า

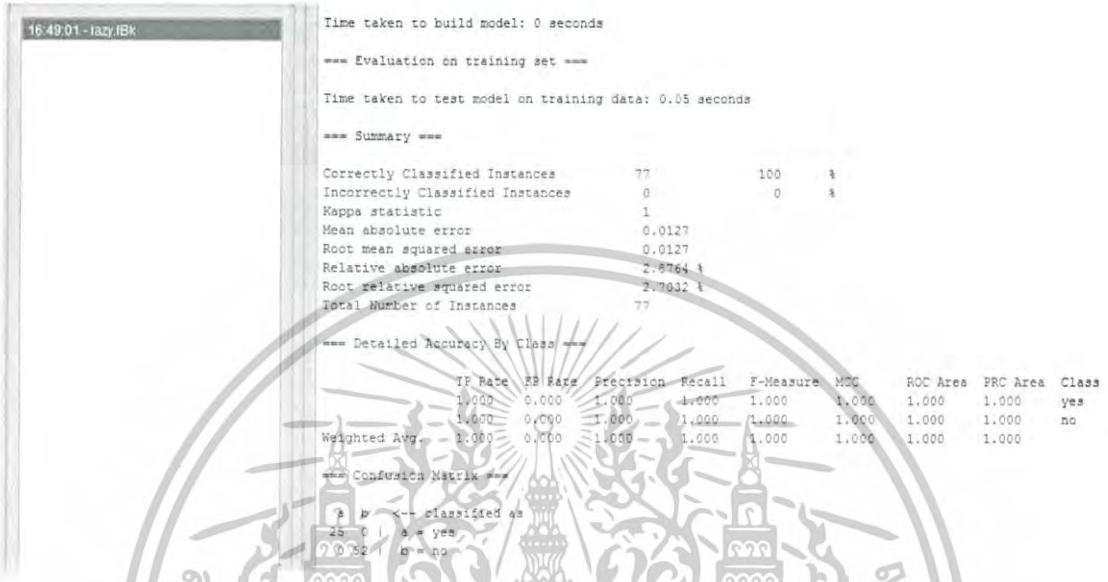
สุ่มครั้งที่ 1 Random Seed = 10 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 70 คน มีค่าทำนายความถูกต้องจำนวน 70 คน คิดเป็น 100% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.0126)^2$ = 0.0002 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0121

สุ่มครั้งที่ 2 Random Seed = 20 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 70 คน มีค่าทำนายความถูกต้องจำนวน 70 คน คิดเป็น 100% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.0123)^2$ = 0.0002 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0117

สุ่มครั้งที่ 3 Random Seed = 30 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 70 คน มีค่าทำนายความถูกต้องจำนวน 70 คน คิดเป็น 100% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.0134)^2$ = 0.0002 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0133

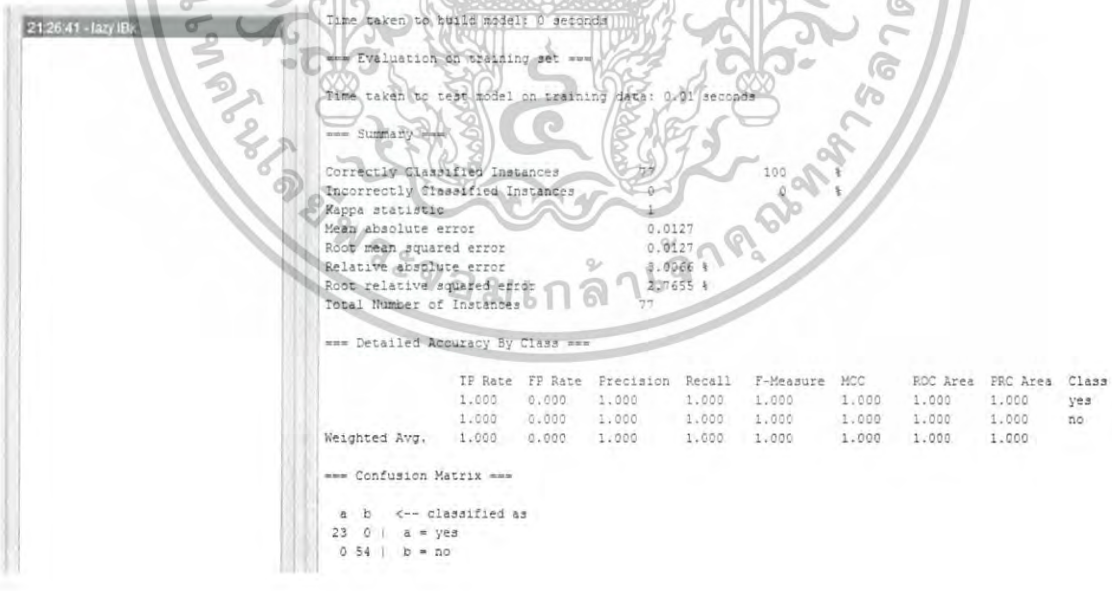
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.3.1.2 โรคเบาหวานของชาวพิม่า ประเทศอินเดีย (Pima Indians Diabetes) จำนวนข้อมูลทดสอบ 768 ค่า ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่ปานกลาง เมื่อกำหนดค่า Random Seed 10 20 และ 30 ตามลำดับ



```
16:49:01 - lazy.IBk

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.05 seconds

=== Summary ===

Correctly Classified Instances          77               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0127
Root mean squared error                  0.0127
Relative absolute error                  2.8764 %
Root relative squared error              2.7032 %
Total Number of Instances               77

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 1.000    0.000      1.000      1.000    1.000      1.000    1.000     1.000     yes
                 1.000    0.000      1.000      1.000    1.000      1.000    1.000     1.000     no
Weighted Avg.    1.000    0.000      1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

  a  b   <-- classified as
 25  0 |  a = yes
  0 52 |  b = no
```

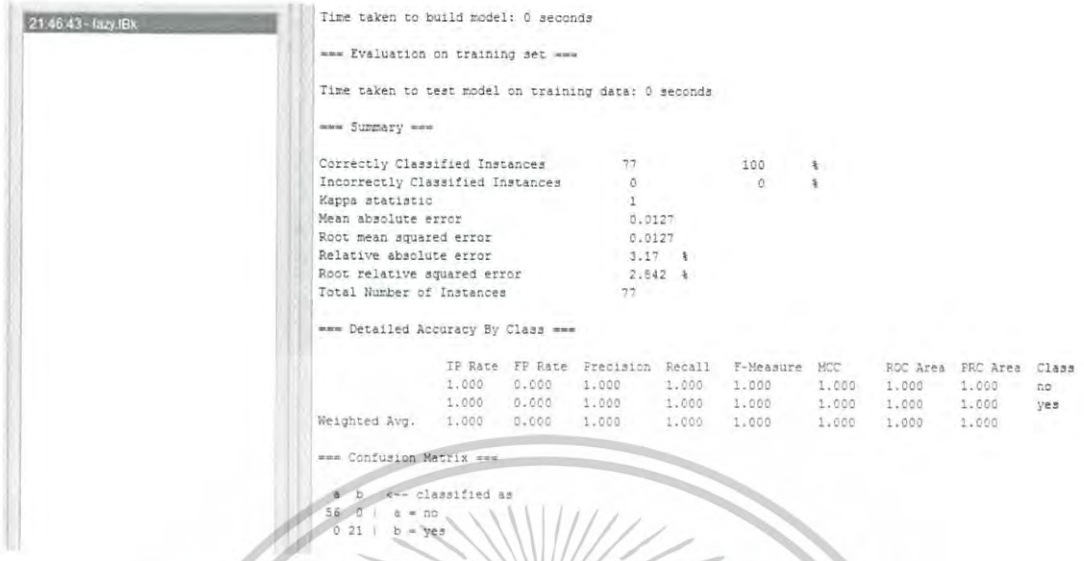
รูปที่ 4-22 ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 10



```
21:26:41 - lazy.IBk

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances          77               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0127
Root mean squared error                  0.0127
Relative absolute error                  3.0966 %
Root relative squared error              2.7655 %
Total Number of Instances               77

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 1.000    0.000      1.000      1.000    1.000      1.000    1.000     1.000     yes
                 1.000    0.000      1.000      1.000    1.000      1.000    1.000     1.000     no
Weighted Avg.    1.000    0.000      1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

  a  b   <-- classified as
 23  0 |  a = yes
  0 54 |  b = no
```

รูปที่ 4-23 ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 20

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4-24 ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 30

จากรูป 4-22 ถึง 4-24 ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK โดยโปรแกรม SPSS โรคเบาหวานของชาวพิม่า ประเทศอินเดีย โดยทำการสุ่ม 3 ครั้ง Random Seed 10 20 และ 30 ตามลำดับ จากผลลัพธ์ที่ได้พบว่า

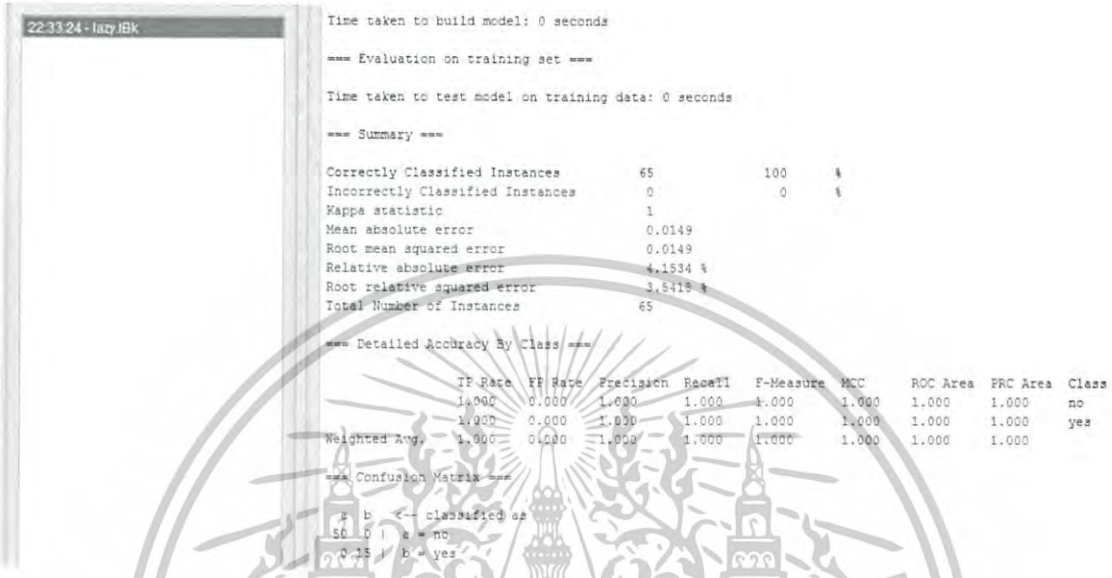
สุ่มครั้งที่ 1 Random Seed = 10 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 77 คน มีค่าทำนายความถูกต้องจำนวน 77 คน คิดเป็น 100% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.0127)^2$ = 0.0002 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0127

สุ่มครั้งที่ 2 Random Seed = 20 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 77 คน มีค่าทำนายความถูกต้องจำนวน 77 คน คิดเป็น 100% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.0127)^2$ = 0.0002 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0127

สุ่มครั้งที่ 3 Random Seed = 30 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 77 คน มีค่าทำนายความถูกต้องจำนวน 77 คน คิดเป็น 100% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.0127)^2$ = 0.0002 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0127

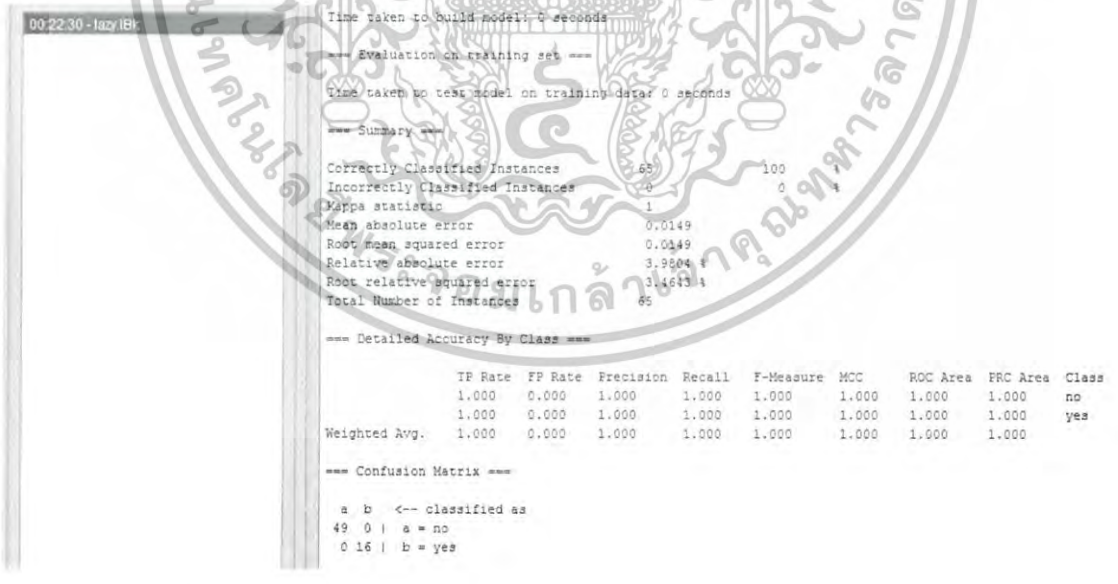
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.3.1.3 การชำระเงินด้วยบัตรเครดิตของลูกค้า (Default of Credit Card Clients) จำนวนข้อมูลทดสอบ 646 ค่า ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่สูง เมื่อกำหนดค่า Random Seed 10 20 และ 30 ตามลำดับ



```
22:33:24 - lazy.IBk

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          65              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0.0149
Root mean squared error                  0.0149
Relative absolute error                  4.1534 %
Root relative squared error              3.5419 %
Total Number of Instances               65

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000      1.000      1.000     1.000      1.000     1.000     1.000     no
                 1.000    0.000      1.000      1.000     1.000      1.000     1.000     1.000     yes
Weighted Avg.    1.000    0.000      1.000      1.000     1.000      1.000     1.000     1.000

=== Confusion Matrix ===

  a  b   <-- classified as
 50  0 |  a = no
  0 15 |  b = yes
```

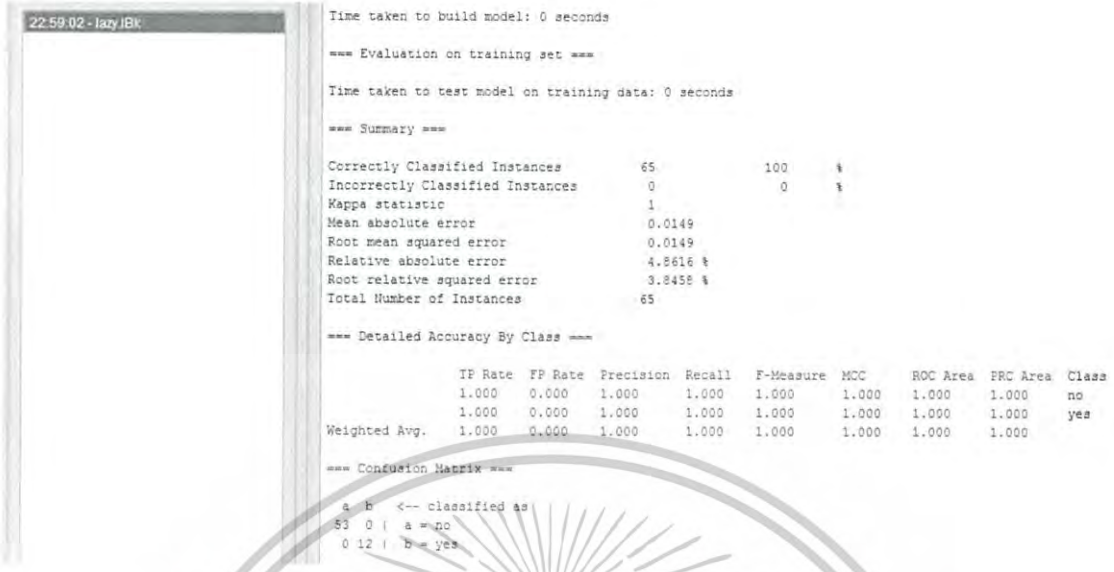
รูปที่ 4-25 ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 10



```
00:22:30 - lazy.IBk

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          65              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0.0149
Root mean squared error                  0.0149
Relative absolute error                  3.9604 %
Root relative squared error              3.4643 %
Total Number of Instances               65

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000      1.000      1.000     1.000      1.000     1.000     1.000     no
                 1.000    0.000      1.000      1.000     1.000      1.000     1.000     1.000     yes
Weighted Avg.    1.000    0.000      1.000      1.000     1.000      1.000     1.000     1.000

=== Confusion Matrix ===

  a  b   <-- classified as
 49  0 |  a = no
  0 16 |  b = yes
```

รูปที่ 4-26 ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 20

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



22:59:02 - lazy.IBk

```
Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          65               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0149
Root mean squared error                  0.0149
Relative absolute error                  4.8616 %
Root relative squared error              3.8458 %
Total Number of Instances               65

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000   1.000      1.000    1.000     1.000     no
                 1.000    0.000    1.000      1.000   1.000      1.000    1.000     1.000     yes
Weighted Avg.    1.000    0.000    1.000      1.000   1.000      1.000    1.000     1.000

=== Confusion Matrix ===

  a  b   <-- classified as
 53  0 |  a = no
  0 12 |  b = yes
```

รูปที่ 4-27 ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 30

จากรูป 4-25 ถึง 4-27 ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK โดยโปรแกรม SPSS การชำระเงินด้วยบัตรเครดิตของลูกค้า โดยทำการสุ่ม 3 ครั้ง Random Seed 10 20 และ 30 ตามลำดับ จากผลลัพธ์ที่ได้พบว่า

สุ่มครั้งที่ 1 Random Seed = 10 กลุ่มลูกค้า 10 เปอร์เซ็นต์ จำนวนทั้งหมด 65 คน มีค่าทำนายความถูกต้องจำนวน 65 คน คิดเป็น 100% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.0149)^2$ = 0.0002 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0149

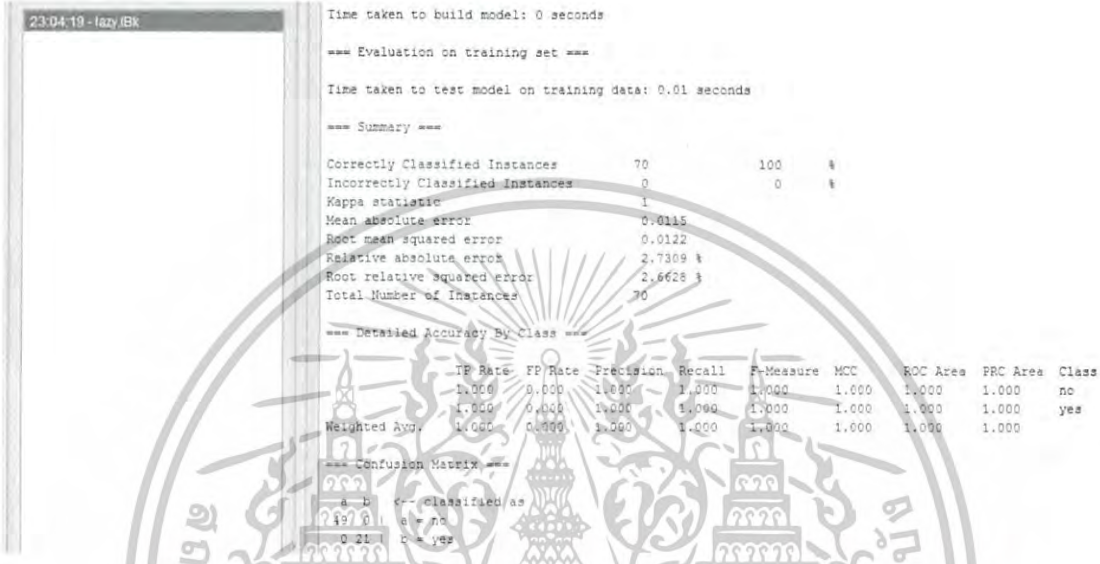
สุ่มครั้งที่ 2 Random Seed = 20 กลุ่มลูกค้า 10 เปอร์เซ็นต์ จำนวนทั้งหมด 65 คน มีค่าทำนายความถูกต้องจำนวน 65 คน คิดเป็น 100% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.0149)^2$ = 0.0002 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0149

สุ่มครั้งที่ 3 Random Seed = 30 กลุ่มลูกค้า 10 เปอร์เซ็นต์ จำนวนทั้งหมด 65 คน มีค่าทำนายความถูกต้องจำนวน 65 คน คิดเป็น 100% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.0149)^2$ = 0.0002 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0149

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

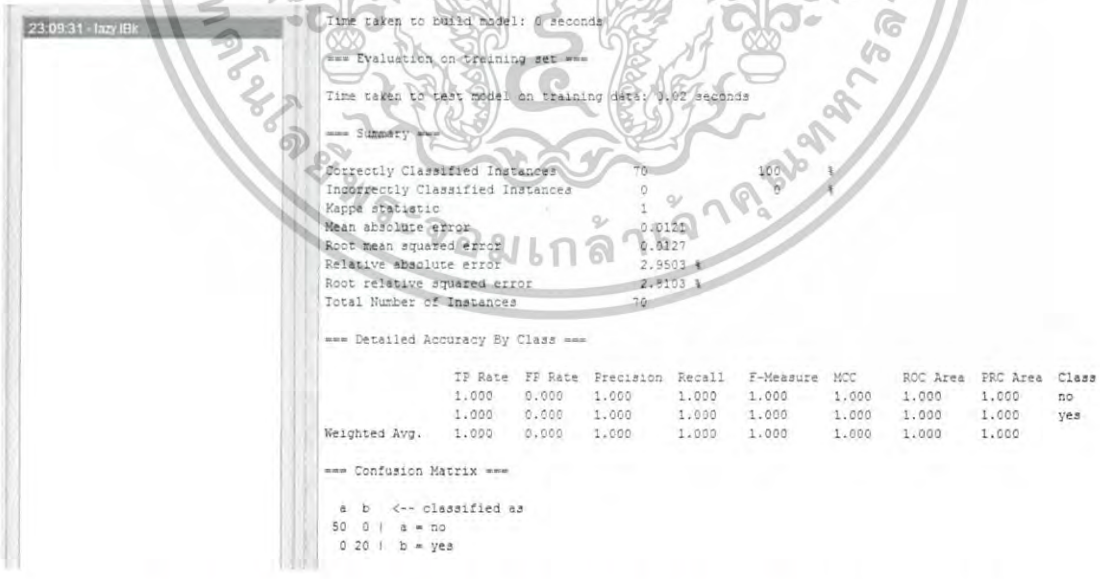
4.3.2 ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBk โดยโปรแกรม WEKA

4.3.2.1 โรคมะเร็งเต้านมของรัฐวิสคอนซิน (Breast Cancer Wisconsin) จำนวนข้อมูลทดสอบ 699 ค่า ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่ต่ำ เมื่อกำหนดค่า Random Seed 10 20 และ 30 ตามลำดับ



```
23:04:19 - lazy.IBk

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances          70               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0115
Root mean squared error                  0.0122
Relative absolute error                  2.7309 %
Root relative squared error              2.6628 %
Total Number of Instances               70

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000     1.000      1.000     1.000     1.000     1.000     1.000     no
                 1.000    0.000     1.000      1.000     1.000     1.000     1.000     1.000     yes
Weighted Avg.    1.000    0.000     1.000      1.000     1.000     1.000     1.000     1.000

=== Confusion Matrix ===

  a  b   <-- classified as
 49  0 |  a = no
  0 21 |  b = yes
```

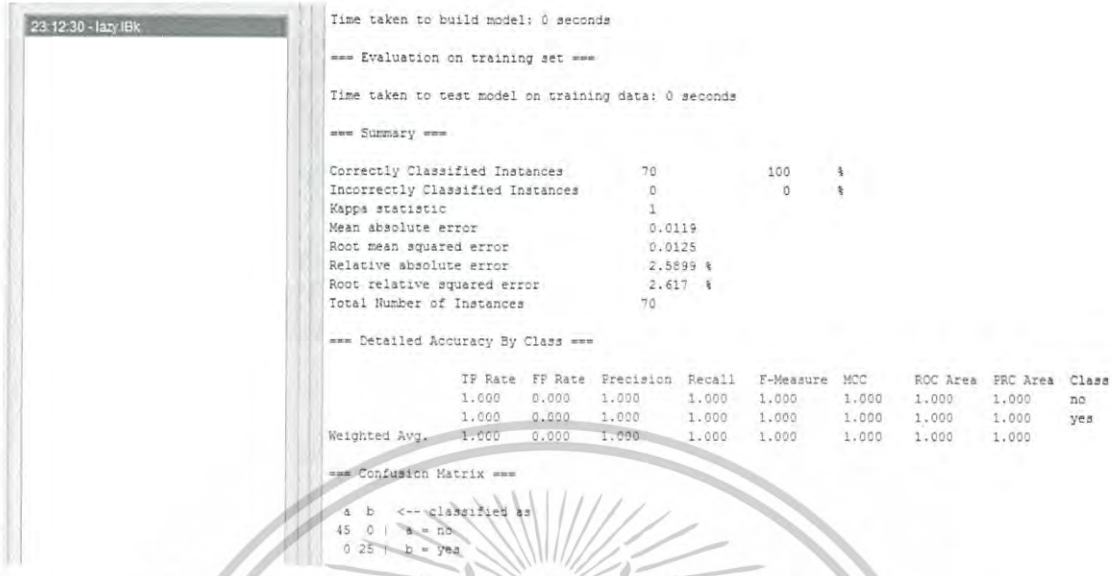
รูปที่ 4-28 ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 10



```
23:09:31 - lazy.IBk

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.02 seconds

=== Summary ===

Correctly Classified Instances          70               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0121
Root mean squared error                  0.0127
Relative absolute error                  2.9503 %
Root relative squared error              2.8103 %
Total Number of Instances               70

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000     1.000      1.000     1.000     1.000     1.000     1.000     no
                 1.000    0.000     1.000      1.000     1.000     1.000     1.000     1.000     yes
Weighted Avg.    1.000    0.000     1.000      1.000     1.000     1.000     1.000     1.000

=== Confusion Matrix ===

  a  b   <-- classified as
 50  0 |  a = no
  0 20 |  b = yes
```

รูปที่ 4-29 ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 20

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4-30 ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 30

จากรูป 4-28 ถึง 4-30 ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK โดยโปรแกรม WEKA โรคมะเร็งเต้านมของรัฐวิสคอนซิน โดยทำการสุ่ม 3 ครั้ง Random Seed 10 20 และ 30 ตามลำดับ จากผลลัพธ์ที่ได้พบว่า

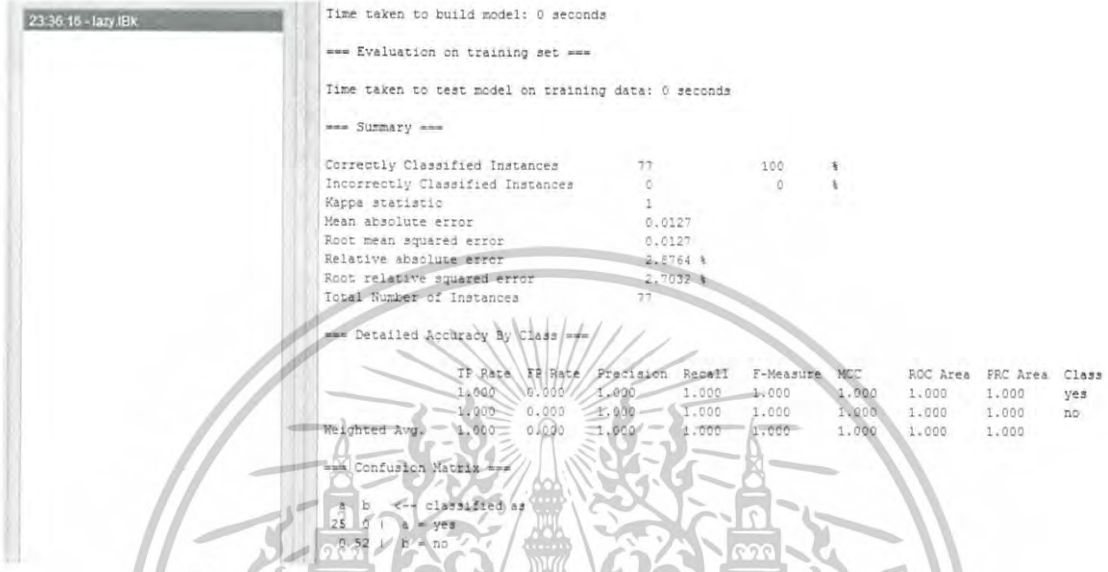
สุ่มครั้งที่ 1 Random Seed = 10 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 70 คน มีค่าทำนายความถูกต้องจำนวน 70 คน คิดเป็น 100% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.0122)^2$ = 0.0005 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0115

สุ่มครั้งที่ 2 Random Seed = 20 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 70 คน มีค่าทำนายความถูกต้องจำนวน 70 คน คิดเป็น 100% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.0127)^2$ = 0.0002 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0121

สุ่มครั้งที่ 3 Random Seed = 30 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 70 คน มีค่าทำนายความถูกต้องจำนวน 70 คน คิดเป็น 100% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.0125)^2$ = 0.0002 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0119

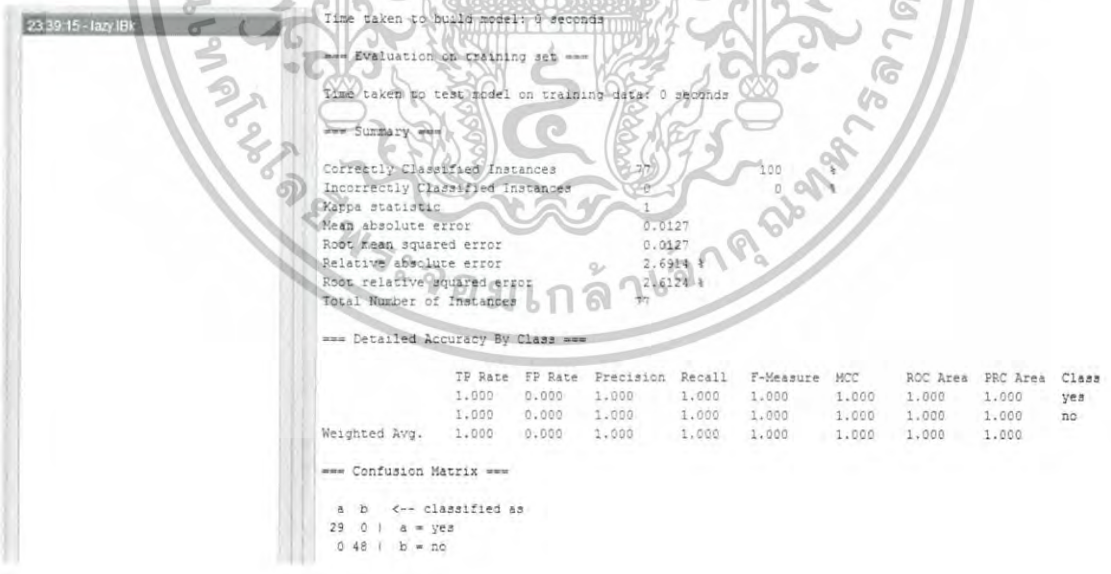
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.3.2.2 โรคเบาหวานของชาวพิม่า ประเทศอินเดีย (Pima Indians Diabetes) จำนวนข้อมูลทดสอบ 768 ค่า ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่ปานกลาง เมื่อกำหนดค่า Random Seed 10 20 และ 30 ตามลำดับ



```
23:36:16 - lazy.IBk

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          77               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0127
Root mean squared error                  0.0127
Relative absolute error                  2.8764 %
Root relative squared error              2.7032 %
Total Number of Instances               77

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               1.000    0.000    1.000      1.000   1.000      1.000    1.000     1.000     yes
               1.000    0.000    1.000      1.000   1.000      1.000    1.000     1.000     no
Weighted Avg.  1.000    0.000    1.000      1.000   1.000      1.000    1.000     1.000

=== Confusion Matrix ===

  a  b   <-- classified as
 25  0 |  a = yes
  0 52 |  b = no
```

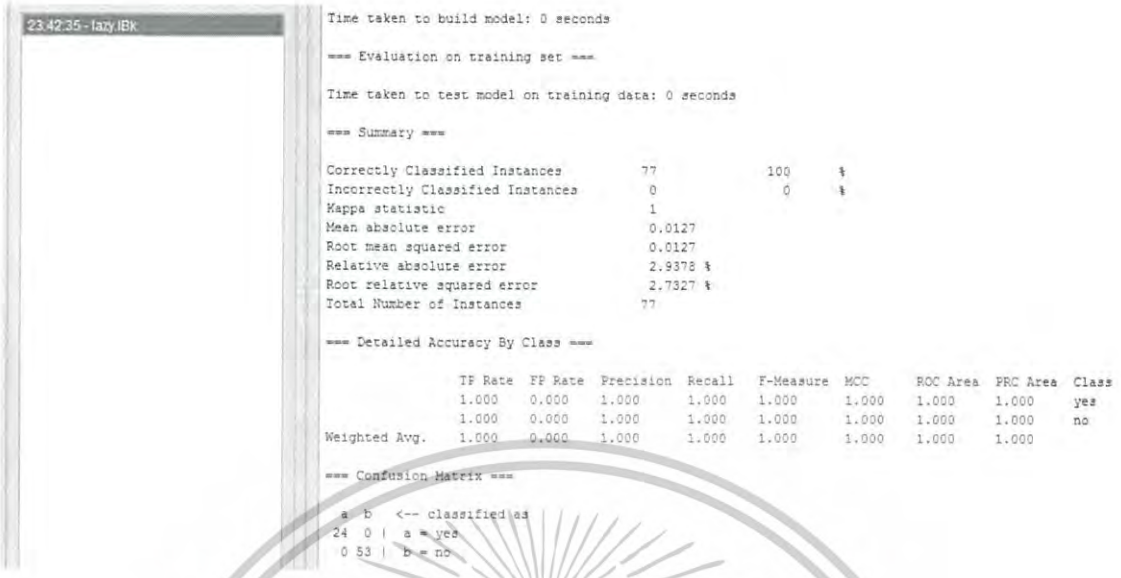
รูปที่ 4-31 ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 10



```
23:39:15 - lazy.IBk

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          77               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0127
Root mean squared error                  0.0127
Relative absolute error                  2.6914 %
Root relative squared error              2.6124 %
Total Number of Instances               77

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               1.000    0.000    1.000      1.000   1.000      1.000    1.000     1.000     yes
               1.000    0.000    1.000      1.000   1.000      1.000    1.000     1.000     no
Weighted Avg.  1.000    0.000    1.000      1.000   1.000      1.000    1.000     1.000

=== Confusion Matrix ===

  a  b   <-- classified as
 29  0 |  a = yes
  0 48 |  b = no
```

รูปที่ 4-32 ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 20

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



23:42:35 - lazy.IBk

```
Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          77               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0127
Root mean squared error                  0.0127
Relative absolute error                  2.9378 %
Root relative squared error              2.7327 %
Total Number of Instances               77

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000   1.000      1.000    1.000     1.000     yes
                 1.000    0.000    1.000      1.000   1.000      1.000    1.000     1.000     no
Weighted Avg.    1.000    0.000    1.000      1.000   1.000      1.000    1.000     1.000

=== Confusion Matrix ===

  a  b   <-- classified as
 24  0 |  a = yes
  0 53 |  b = no
```

รูปที่ 4-33 ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 30

จากรูป 4-31 ถึง 4-33 ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK โดยโปรแกรม WEKA โรคเบาหวานของชาวพิม่า ประเทศอินเดีย โดยทำการสุ่ม 3 ครั้ง Random Seed 10 20 และ 30 ตามลำดับ จากผลลัพธ์ที่ได้พบว่า

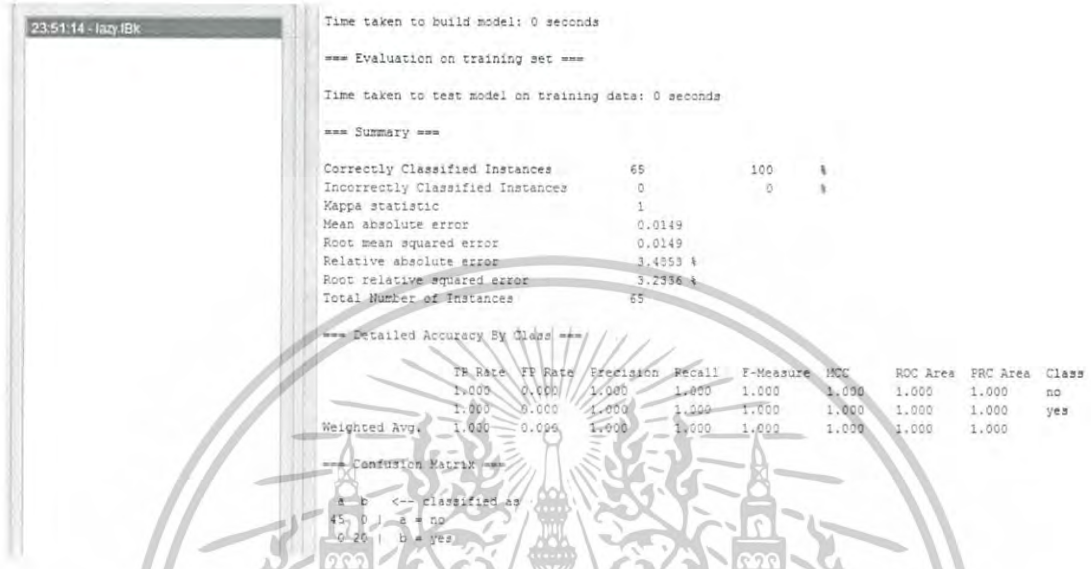
สุ่มครั้งที่ 1 Random Seed = 10 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 77 คน มีค่าทำนายความถูกต้องจำนวน 77 คน คิดเป็น 100% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.0127)^2$ = 0.0002 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0127

สุ่มครั้งที่ 2 Random Seed = 20 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 77 คน มีค่าทำนายความถูกต้องจำนวน 77 คน คิดเป็น 100% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.0127)^2$ = 0.0002 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0127

สุ่มครั้งที่ 3 Random Seed = 30 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 77 คน มีค่าทำนายความถูกต้องจำนวน 77 คน คิดเป็น 100% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.0127)^2$ = 0.0002 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0127

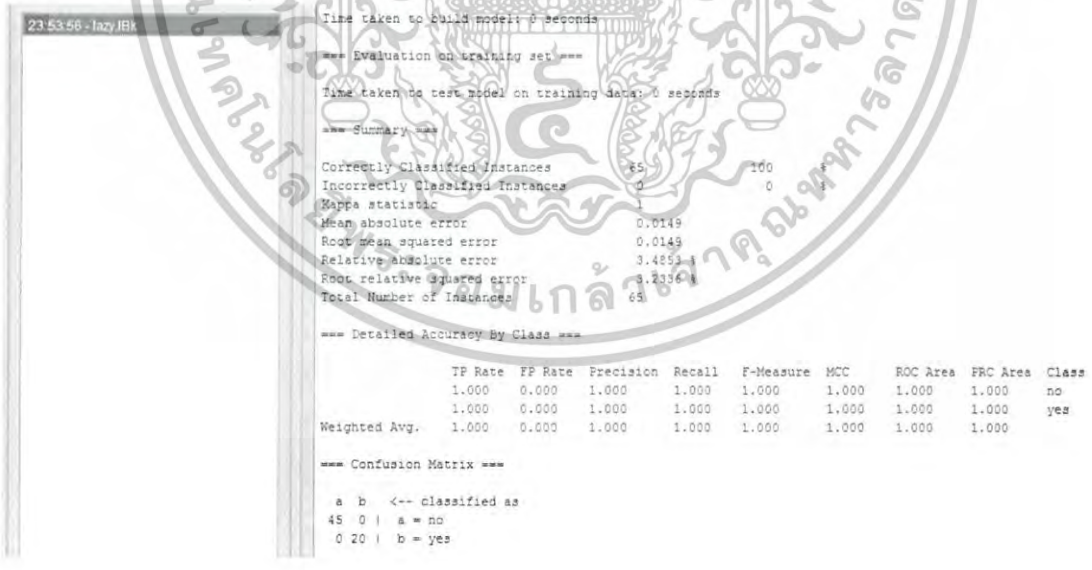
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.3.2.3 การชำระเงินด้วยบัตรเครดิตของลูกค้า (Default of Credit Card Clients) จำนวนข้อมูลทดสอบ 646 ค่า ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่สูง เมื่อกำหนดค่า Random Seed 10 20 และ 30 ตามลำดับ



```
23:51:14 - lazy.IBk

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          65               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0149
Root mean squared error                  0.0149
Relative absolute error                  3.4853 %
Root relative squared error              3.2336 %
Total Number of Instances               65

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000      1.000     1.000     1.000     1.000     1.000     1.000     no
                 1.000    0.000      1.000     1.000     1.000     1.000     1.000     1.000     yes
Weighted Avg.    1.000    0.000      1.000     1.000     1.000     1.000     1.000     1.000

=== Confusion Matrix ===

  a  b   <-- classified as
 45  0 |  a = no
  0 20 |  b = yes
```

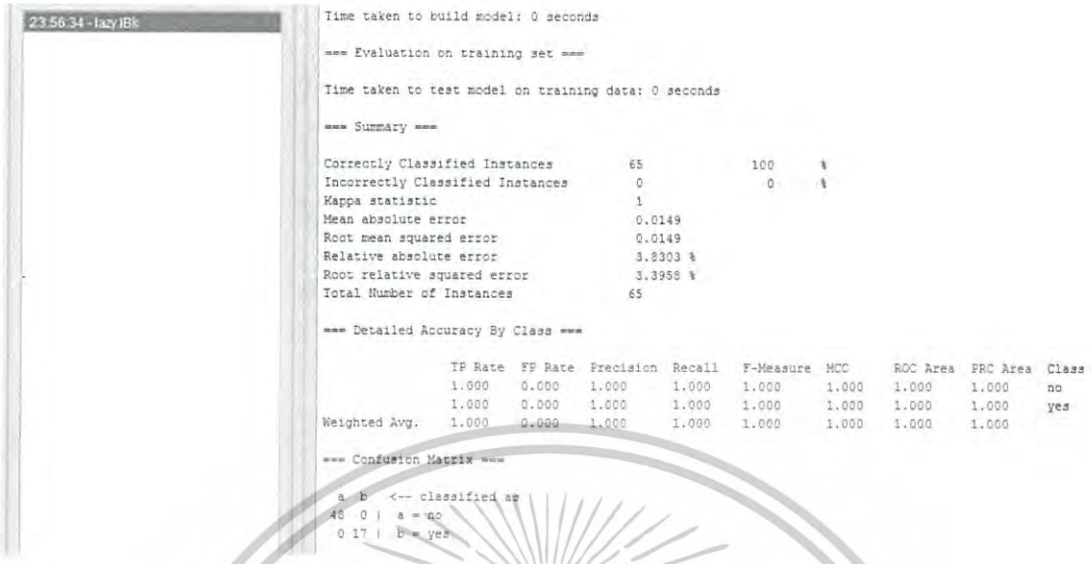
รูปที่ 4-34 ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 10



```
23:53:56 - lazy.IBk

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          65               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0149
Root mean squared error                  0.0149
Relative absolute error                  3.4853 %
Root relative squared error              3.2336 %
Total Number of Instances               65

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000      1.000     1.000     1.000     1.000     1.000     1.000     no
                 1.000    0.000      1.000     1.000     1.000     1.000     1.000     1.000     yes
Weighted Avg.    1.000    0.000      1.000     1.000     1.000     1.000     1.000     1.000

=== Confusion Matrix ===

  a  b   <-- classified as
 45  0 |  a = no
  0 20 |  b = yes
```

รูปที่ 4-35 ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 20

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



23:56:34 - lazy.IBk

```
Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          65               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0149
Root mean squared error                  0.0149
Relative absolute error                  3.8303 %
Root relative squared error              3.3958 %
Total Number of Instances               65

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     no
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     yes
Weighted Avg.  1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

  a  b   <-- classified as
 48  0 |  a = no
  0 17 |  b = yes
```

รูปที่ 4-36 ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 30

จากรูป 4-34 ถึง 4-36 ผลการวิเคราะห์ข้อมูลด้วยวิธีเพื่อนบ้านใกล้สุด k ตัวโดยใช้อัลกอริทึม IBK โดยโปรแกรม WEKA การชำระเงินด้วยบัตรเครดิตของลูกค้า โดยทำการสุ่ม 3 ครั้ง Random Seed 10 20 และ 30 ตามลำดับ จากผลลัพธ์ที่ได้พบว่า

สุ่มครั้งที่ 1 Random Seed = 10 กลุ่มลูกค้า 10 เปอร์เซ็นต์ จำนวนทั้งหมด 65 คน มีค่าทำนายความถูกต้องจำนวน 65 คน คิดเป็น 100% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.0149)^2$ = 0.0002 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0149

สุ่มครั้งที่ 2 Random Seed = 20 กลุ่มลูกค้า 10 เปอร์เซ็นต์ จำนวนทั้งหมด 65 คน มีค่าทำนายความถูกต้องจำนวน 65 คน คิดเป็น 100% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.0149)^2$ = 0.0002 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0149

สุ่มครั้งที่ 3 Random Seed = 30 กลุ่มลูกค้า 10 เปอร์เซ็นต์ จำนวนทั้งหมด 65 คน มีค่าทำนายความถูกต้องจำนวน 65 คน คิดเป็น 100% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.0149)^2$ = 0.0002 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0149

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.4 ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48

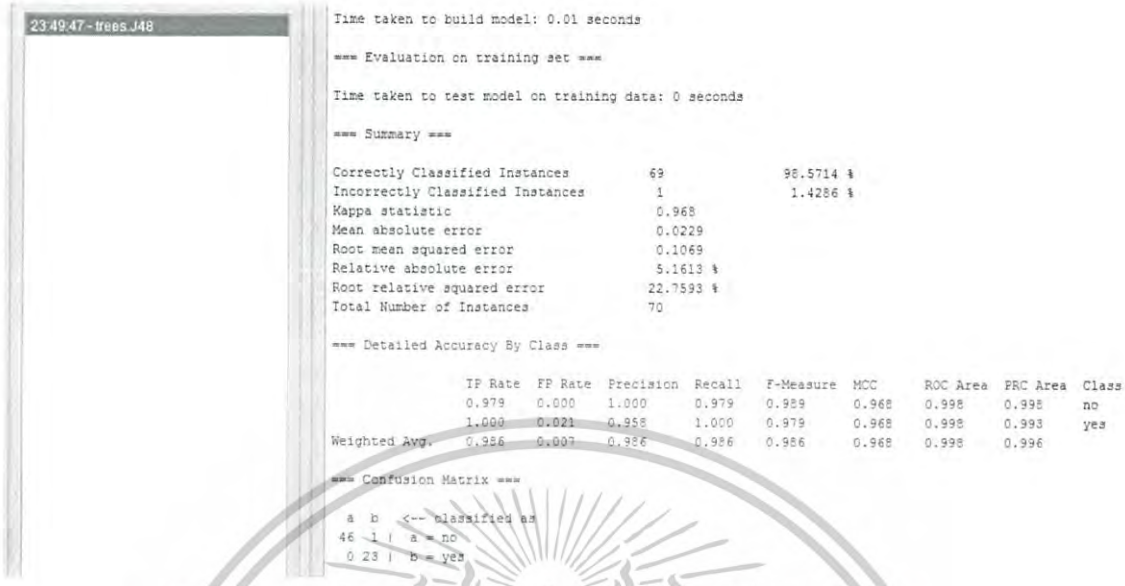
4.4.1 ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 โดยโปรแกรม SPSS

4.4.1.1 โรคมะเร็งเต้านมของรัฐวิสคอนซิน (Breast Cancer Wisconsin) จำนวนข้อมูลทดสอบ 699 ค่า ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่ต่ำ เมื่อกำหนดค่า Random Seed 10 20 และ 30 ตามลำดับ

รูปที่ 4-37 ต้นไม้เพื่อการตัดสินใจ โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 10

จากรูป 4-37 แสดงต้นไม้เพื่อการตัดสินใจ โรคมะเร็งเต้านมของรัฐวิสคอนซิน โหนดรากแบ่งโดยตัวแปรทำนายคือ X_3 = ความสม่ำเสมอของรูปร่างเซลล์ ถ้าผลน้อยกว่าเท่ากับ 2 ยูนิตเซลล์ คือไม่เป็นโรคมะเร็งเต้านม (จำนวน 43 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_7 = ของเหลวในเซลล์ ถ้าผลมากกว่า 3 นาโนเมตร คือ เป็นโรคมะเร็งเต้านม (จำนวน 19 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_1 = ความหนาของก้อนเนื้อ ผลน้อยกว่าเท่ากับ 6 มิลลิเมตร คือ ไม่เป็นโรคมะเร็งเต้านม (จำนวน 3 คน) ถ้าผลมากกว่า 6 มิลลิเมตร คือ เป็นโรคมะเร็งเต้านม (จำนวน 5 คน)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

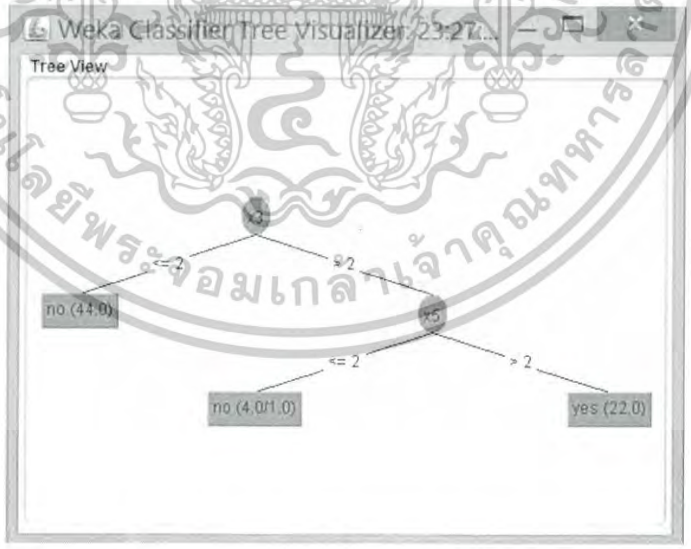


23:49:47 - trees.J48

```
Time taken to build model: 0.01 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          69               98.5714 %
Incorrectly Classified Instances         1                1.4286 %
Kappa statistic                          0.968
Mean absolute error                      0.0229
Root mean squared error                  0.1069
Relative absolute error                  5.1613 %
Root relative squared error             22.7593 %
Total Number of Instances               70

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.979    0.000    1.000      0.979   0.989      0.968    0.998     0.998     no
               1.000    0.021    0.958      1.000   0.979      0.968    0.998     0.993     yes
Weighted Avg.  0.986    0.007    0.986      0.986   0.986      0.968    0.998     0.996

=== Confusion Matrix ===

  a  b   <-- classified as
 46  1 |  a = no
  0 23 |  b = yes
```

รูปที่ 4-38 ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 10

จากรูป 4-38 สุ่มครั้งที่ 1 Random Seed = 10 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 70 คน มีค่าทำนายความถูกต้องจำนวน 69 คน คิดเป็น 98.5714% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.1069)^2$ = 0.0114 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0229

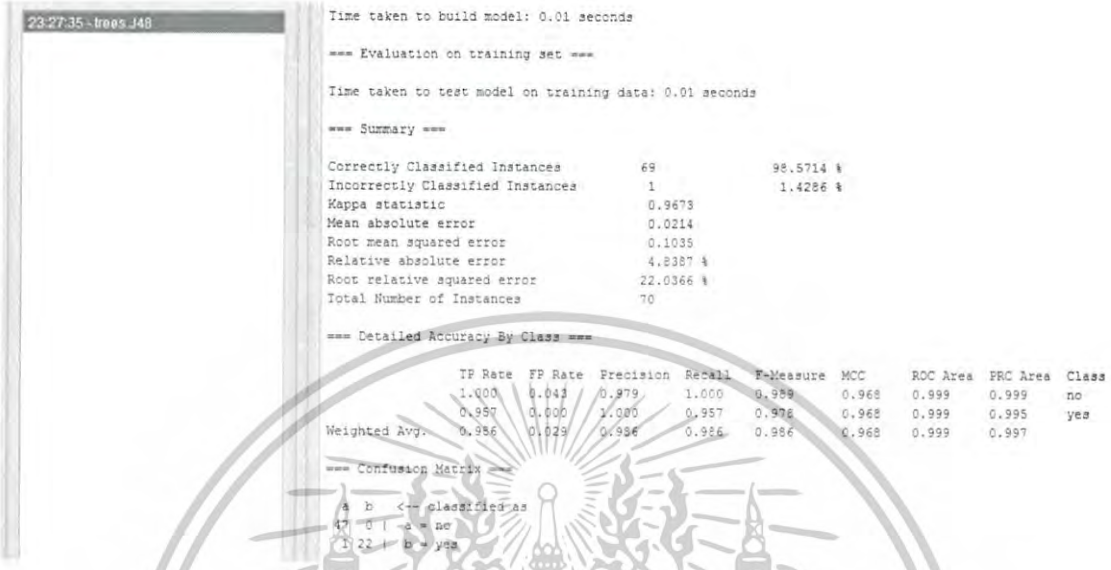


รูปที่ 4-39 ต้นไม้เพื่อการตัดสินใจ โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 20

จากรูป 4-39 แสดงต้นไม้เพื่อการตัดสินใจ โรคมะเร็งเต้านมของรัฐวิสคอนซิน โหนดรากแบ่งโดยตัวแปรทำนายคือ X3 = ความสม่ำเสมอของรูปร่างเซลล์ ถ้าผลน้อยกว่าเท่ากับ 2 ยูนิตเซลล์ คือไม่เป็นโรคมะเร็งเต้านม (จำนวน 44 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_5 = ขนาดเซลล์เดียว ถ้าผล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเอกสารทุกครั้งที่มีการนำไปใช้

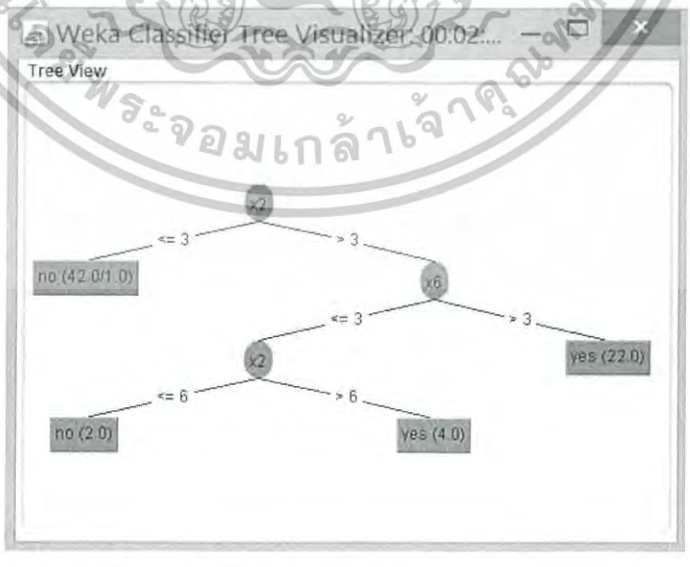
น้อยกว่าเท่ากับ 2 มิลลิเมตร คือ ไม่เป็นโรคมะเร็งเต้านม (จำนวน 4 คน) ถ้าผลมากกว่า 2 มิลลิเมตร คือ เป็นโรคมะเร็งเต้านม (จำนวน 22 คน)



```
23:27:35 - trees.J48

Time taken to build model: 0.01 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances          69               98.5714 %
Incorrectly Classified Instances         1                1.4286 %
Kappa statistic                          0.9673
Mean absolute error                      0.0214
Root mean squared error                  0.1035
Relative absolute error                  4.8387 %
Root relative squared error             22.0366 %
Total Number of Instances               70

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.043    0.979      1.000   0.989      0.968    0.999     0.999     no
                 0.957    0.000    1.000      0.957   0.978      0.968    0.999     0.995     yes
Weighted Avg.    0.986    0.029    0.986      0.986   0.986      0.968    0.999     0.997

=== Confusion Matrix ===

  a  b   <-- classified as
 47  0 |  a = no
  1 22 |  b = yes
```

รูปที่ 4-40 ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 20

จากรูป 4-40 สุ่มครั้งที่ 2 Random Seed = 20 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 70 คน มีค่าทำนายความถูกต้องจำนวน 69 คน คิดเป็น 98.5714% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.1035)^2$ = 0.01071 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0214



รูปที่ 4-41 ต้นไม้เพื่อการตัดสินใจ โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 30

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากรูป 4-41 แสดงต้นไม้เพื่อการตัดสินใจ โรคมะเร็งเต้านมของรัฐวิสคอนซิน โหนดรากแบ่งโดยตัวแปรทำนายคือ X_2 = ความสม่ำเสมอของขนาดเซลล์ ถ้าผลน้อยกว่าเท่ากับ 3 นาโนเมตร คือไม่เป็นโรคมะเร็งเต้านม (จำนวน 42 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_6 = นิวเคลียสไม่ถูกห่อหุ้ม ถ้าผลน้อยกว่าเท่ากับ 3 นิวเคลียส คือ เป็นโรคมะเร็งเต้านม (จำนวน 22 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_2 = ความสม่ำเสมอของขนาดเซลล์ ถ้าผลน้อยกว่าเท่า 6 นาโนเมตร คือไม่เป็นโรคมะเร็งเต้านม (จำนวน 2 คน) ผลมากกว่า 6 นาโนเมตร คือ เป็นโรคมะเร็งเต้านม (จำนวน 4 คน)

00:02:41 - trees.J48

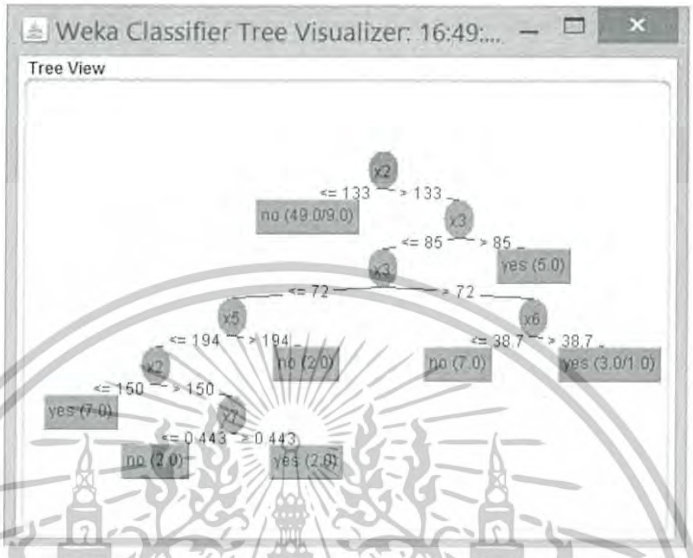
```
Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          69               98.5714 %
Incorrectly Classified Instances         1                1.4286 %
Kappa statistic                          0.9696
Mean absolute error                      0.0279
Root mean squared error                  0.1181
Relative absolute error                  5.8767 %
Root relative squared error             24.26   %
Total Number of Instances               70

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 1.000    0.037      0.977      1.000    0.989      0.970    0.982     0.978     no
                 0.963    0.000      1.000      0.963    0.981      0.970    0.982     0.978     yes
Weighted Avg.    0.986    0.023      0.986      0.986    0.986      0.970    0.982     0.978

=== Confusion Matrix ===

  a  b   <-- classified as
 43  0 |  a = no
  1 26 |  b = yes
```

รูปที่ 4-42 ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 30

จากรูป 4-42 สุ่มครั้งที่ 3 Random Seed = 30 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 70 คน มีค่าทำนายความถูกต้องจำนวน 69 คน คิดเป็น 98.5714% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.1181)^2$ = 0.01395 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.027

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

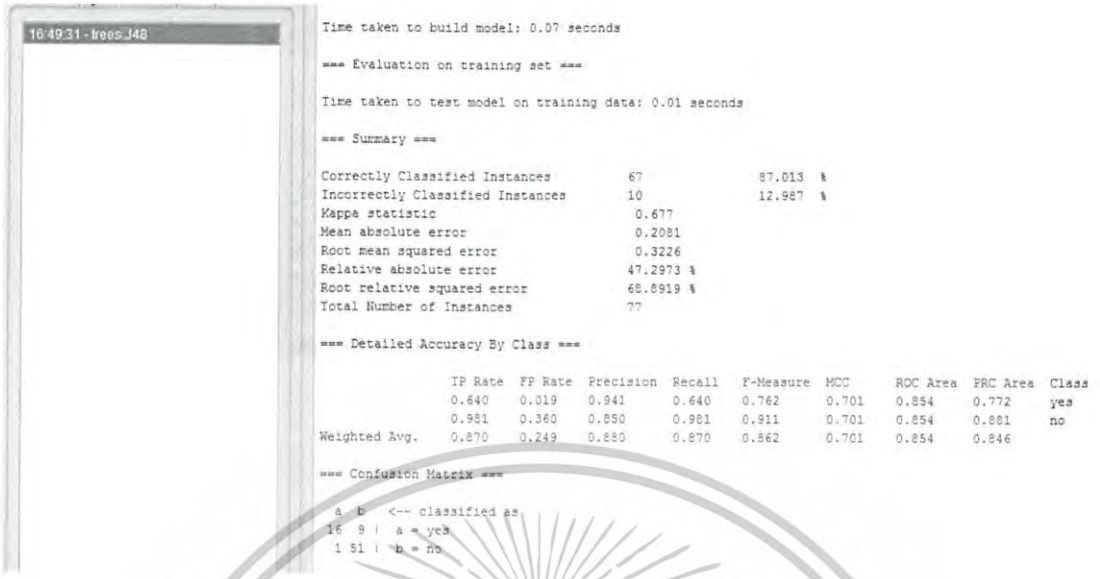
4.4.1.2 โรคเบาหวานของชาวพิม่า ประเทศอินเดีย (Pima Indians Diabetes) จำนวนข้อมูลทดสอบ 768 ค่า ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่ปานกลาง เมื่อกำหนดค่า Random Seed 10 20 และ 30 ตามลำดับ



รูปที่ 4-43 ต้นไม้เพื่อการตัดสินใจ โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 10

จากรูป 4-43 แสดงต้นไม้เพื่อการตัดสินใจ โรคเบาหวานของชาวพิม่า ประเทศอินเดียโหนดรากแบ่งโดยตัวแปรทำนายคือ X_2 = ความเข้มข้นของกลูโคสในเลือด 2 ชั่วโมงก่อนการทดสอบระดับกลูโคสในช่องปาก ถ้าผลน้อยกว่าเท่ากับ 133 มิลลิกรัมต่อเดซิเบล คือไม่เป็นโรคเบาหวาน (จำนวน 49 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_3 = ความดันโลหิตสูง ผลมากกว่า 85 มิลลิเมตรปรอท คือ เป็นโรคเบาหวาน (จำนวน 5 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_6 = ค่าดัชนีมวลกาย ถ้าผลน้อยกว่าเท่ากับ 38.7 กิโลกรัมต่อตารางเมตร คือ ไม่เป็นเบาหวาน (จำนวน 7 คน) ผลมากกว่า 38.7 กิโลกรัมต่อตารางเมตร คือ เป็นโรคเบาหวาน (จำนวน 3 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_5 = จำนวนอินซูลิน หลังจากอดอาหารภายใน 2 ชั่วโมง ถ้าผลมากกว่า 194 ล้านยูนิตต่อมิลลิลิตร คือ ไม่เป็นโรคเบาหวาน (จำนวน 2 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_2 = ความเข้มข้นของกลูโคสในเลือด 2 ชั่วโมงก่อนการทดสอบระดับกลูโคสในช่องปาก ถ้าผลน้อยกว่าเท่ากับ 150 มิลลิกรัมต่อเดซิเบล คือ เป็นโรคเบาหวาน (จำนวน 7 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_7 = พันธุกรรมของผู้ป่วยที่มีความเสี่ยงเป็นโรคเบาหวาน ถ้าผลน้อยกว่าเท่ากับร้อยละ 0.443 คือ ไม่เป็นโรคเบาหวาน (จำนวน 2 คน) ถ้าผลมากกว่าร้อยละ 0.443 คือ เป็นโรคเบาหวาน (จำนวน 2 คน)

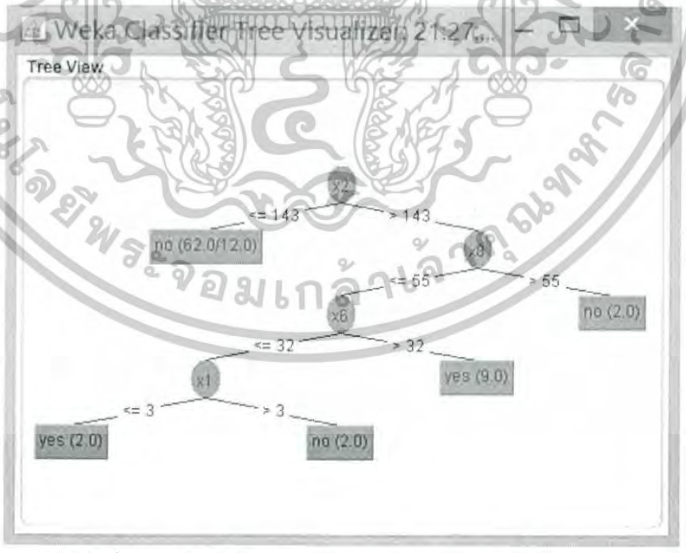
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



```
16:49:31 - trees.J48

Time taken to build model: 0.07 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances          67               87.013 %
Incorrectly Classified Instances        10               12.987 %
Kappa statistic                          0.677
Mean absolute error                      0.2081
Root mean squared error                  0.3226
Relative absolute error                 47.2973 %
Root relative squared error             68.8919 %
Total Number of Instances               77

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.640    0.019    0.941      0.640   0.762      0.701    0.854     0.772     yes
               0.981    0.360    0.850      0.981   0.911      0.701    0.854     0.881     no
Weighted Avg.  0.870    0.249    0.880      0.870   0.862      0.701    0.854     0.846

=== Confusion Matrix ===

  a  b   <-- classified as
 16  9 |  a = yes
  1 51 |  b = no
```

รูปที่ 4-44 ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 10

จากรูป 4-44 สุ่มครั้งที่ 1 Random Seed = 10 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 77 คน มีค่าทำนายความถูกต้องจำนวน 67 คน คิดเป็น 87.013% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.3226)^2$ = 0.1041 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.2081

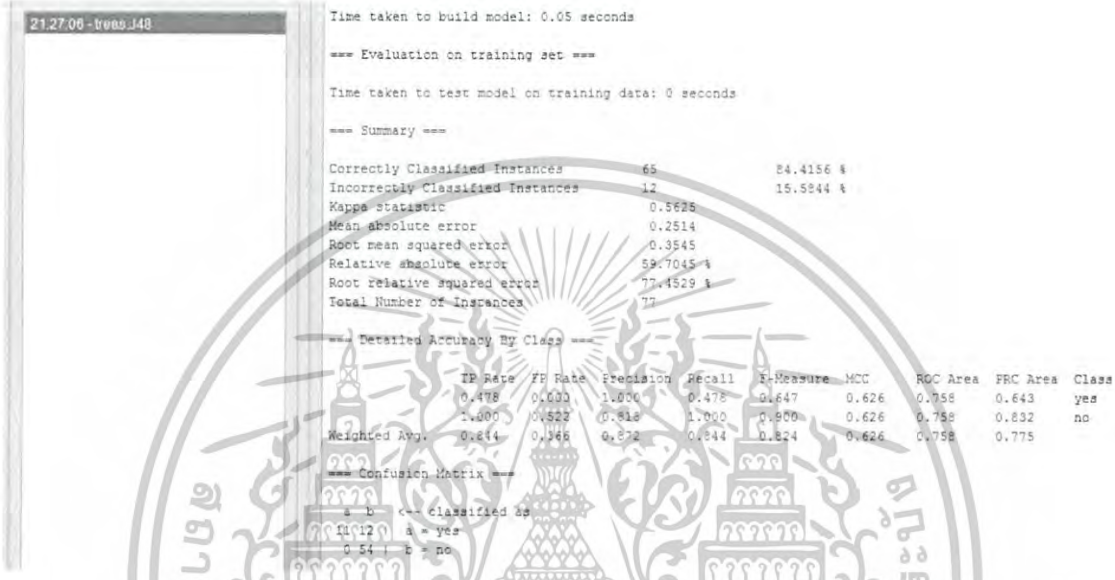


รูปที่ 4-45 ต้นไม้เพื่อการตัดสินใจ โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 20

จากรูป 4-45 แสดงต้นไม้เพื่อการตัดสินใจ โรคเบาหวานของชาวพิม่า ประเทศอินเดียโหนดรากแบ่งโดยตัวแปรทำนายคือ X_2 = ความเข้มข้นของกลูโคสในเลือด 2 ชั่วโมงก่อนการทดสอบระดับกลูโคสในช่องปาก ถ้าผลน้อยกว่าเท่ากับ 143 มิลลิกรัมต่อเดซิเบล คือไม่เป็นโรคเบาหวาน (จำนวน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

62 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_8 = อายุ ถ้าผลมากกว่า 55 ปี คือ ไม่เป็นโรคเบาหวาน (จำนวน 2 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_6 = ค่าดัชนีมวลกาย ถ้าผลมากกว่า 32 กิโลกรัมต่อตารางเมตร คือ เป็นโรคเบาหวาน (จำนวน 9 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_1 = จำนวนครั้งของหญิงที่ตั้งครรภ์มาแล้ว ถ้าผลน้อยกว่าเท่ากับ 3 ครั้ง คือ เป็นโรคเบาหวาน (จำนวน 2 คน) ผลมากกว่า 2 ครั้ง คือ ไม่เป็นโรคเบาหวาน (จำนวน 2 คน)



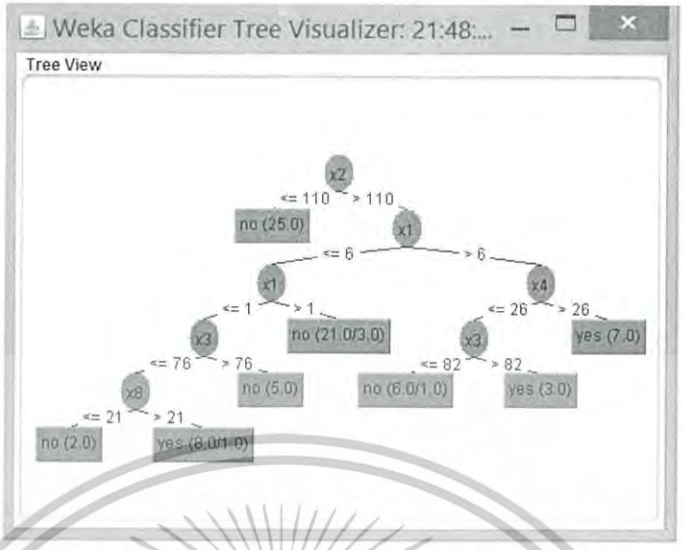
21:27:06 - trees.J48

```
Time taken to build model: 0.05 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          65               84.4156 %
Incorrectly Classified Instances        12               15.5844 %
Kappa statistic                          0.5625
Mean absolute error                      0.2514
Root mean squared error                  0.3545
Relative absolute error                 59.7045 %
Root relative squared error             77.4529 %
Total Number of Instances               77

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.478     0.000       1.000     0.478     0.647     0.626     0.758     0.643     yes
                 1.000     0.522       0.818     1.000     0.900     0.626     0.758     0.832     no
Weighted Avg.    0.844     0.366       0.872     0.844     0.824     0.626     0.758     0.775

=== Confusion Matrix ===

  a  b   <-- classified as
 11 12 |  a = yes
  0 54 |  b = no
```

รูปที่ 4-46 ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 20

จากรูป 4-46 สุ่มครั้งที่ 2 Random Seed = 20 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 77 คน มีค่าทำนายความถูกต้องจำนวน 65 คน คิดเป็น 84.4156% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.3545)^2$ = 0.1257 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.251

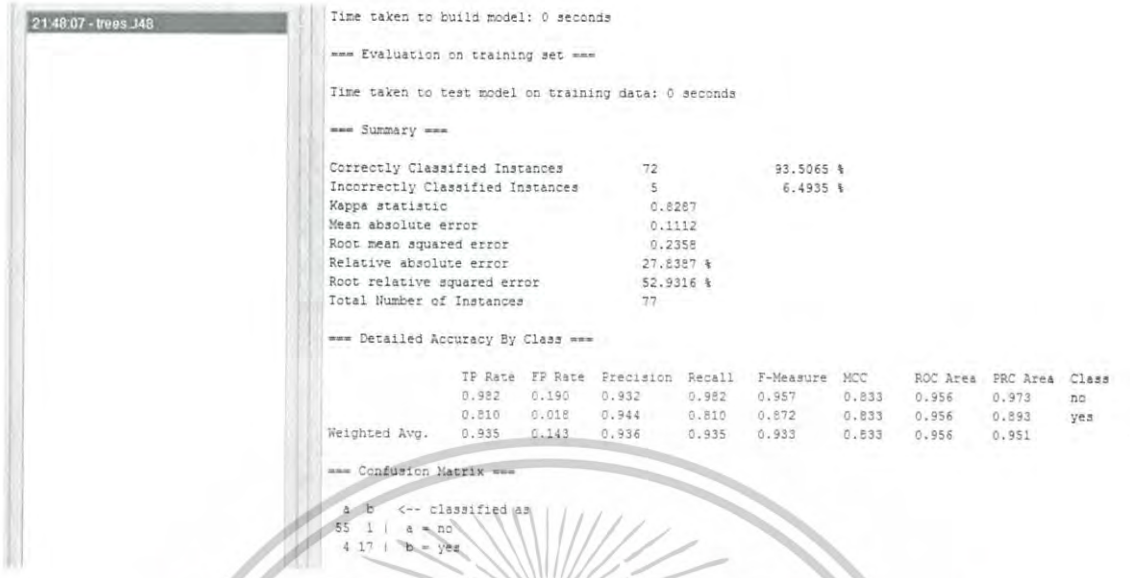
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4-47 ต้นไม้เพื่อการตัดสินใจ โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 30

จากรูป 4-47 แสดงต้นไม้เพื่อการตัดสินใจ โรคเบาหวานของชาวพิม่า ประเทศอินเดียโหนดรากแบ่งโดยตัวแปรทำนายคือ X_2 = ความเข้มข้นของกลูโคสในเลือด 2 ชั่วโมงก่อนการทดสอบระดับกลูโคสในช่องปาก ถ้าผลน้อยกว่าเท่ากับ 110 มิลลิกรัมต่อเดซิเบล คือไม่เป็นโรคเบาหวาน (จำนวน 25 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_1 = จำนวนครั้งของหญิงที่ตั้งครรภ์มาแล้ว ผลมากกว่า 1 ครั้ง คือ ไม่เป็นโรคเบาหวาน (จำนวน 21 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_3 = ความดันโลหิตสูง ถ้าผลมากกว่า 76 มิลลิเมตรปรอท คือ ไม่เป็นโรคเบาหวาน (จำนวน 5 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_8 = อายุ ถ้าผลน้อยกว่าเท่ากับ 21 ปี คือ ไม่เป็นโรคเบาหวาน (จำนวน 2 คน) ผลมากกว่า 21 ปี คือ เป็นโรคเบาหวาน (จำนวน 8 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_4 = ความหนาของชั้นไขมันใต้ผิวหนังบริเวณกล้ามที่ด้านหน้าของต้นแขน ถ้าผลมากกว่า 26 มิลลิเมตร คือ เป็นโรคเบาหวาน (จำนวน 7 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_3 = ความดันโลหิตสูง ถ้าผลน้อยกว่าเท่ากับ 82 มิลลิเมตรปรอท คือ ไม่เป็นโรคเบาหวาน (จำนวน 6 คน) ผลมากกว่า 82 มิลลิเมตรปรอท คือ เป็นโรคเบาหวาน (จำนวน 3 คน)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



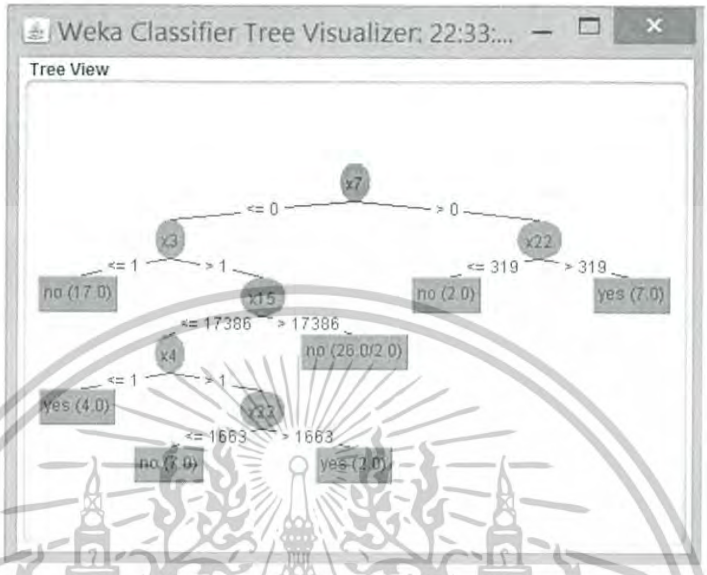
```
21:48:07 - trees.J48

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          72               93.5065 %
Incorrectly Classified Instances         5                6.4935 %
Kappa statistic                          0.8287
Mean absolute error                      0.1112
Root mean squared error                  0.2358
Relative absolute error                 27.8387 %
Root relative squared error             52.9316 %
Total Number of Instances               77

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.982    0.190    0.932      0.982   0.957      0.833    0.956     0.973     no
                 0.810    0.018    0.944      0.810   0.872      0.833    0.956     0.893     yes
Weighted Avg.    0.935    0.143    0.936      0.935   0.933      0.833    0.956     0.951

=== Confusion Matrix ===

  a  b   <-- classified as
 55  1 |  a = no
  4 17 |  b = yes
```

รูปที่ 4-48 ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 30

จากรูป 4-48 สุ่มครั้งที่ 3 Random Seed = 30 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 77 คน มีค่าทำนายความถูกต้องจำนวน 72 คน คิดเป็น 93.5065% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.2358)^2$ = 0.0556 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.1112

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.4.1.3 การชำระเงินด้วยบัตรเครดิตของลูกค้า (Default of Credit Card Clients) จำนวนข้อมูลทดสอบ 646 ค่า ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่สูง เมื่อกำหนดค่า Random Seed 10 20 และ 30 ตามลำดับ



รูปที่ 4-49 ต้นไม้เพื่อการตัดสินใจ การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 10

จากรูป 4-49 แสดงต้นไม้เพื่อการตัดสินใจ การชำระเงินด้วยบัตรเครดิตของลูกค้า โหนดรากแบ่งโดยตัวแปรทำนายคือ X_7 = สถานการณ์ชำระเงินในเดือนสิงหาคม ตัดสินใจต่อด้วยตัวแปรทำนาย X_{22} = จำนวนเงินที่จ่ายในเดือนพฤษภาคม ถ้าผลน้อยกว่าเท่ากับ 319 บาท คือ ไม่ชำระเงินค่าบัตรเครดิต (จำนวน 2 คน) ผลมากกว่า 319 บาท คือ ชำระเงินค่าบัตรเครดิต (จำนวน 7 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_3 = การศึกษา ถ้าผลสำเร็จการศึกษาแล้ว คือ ไม่ชำระเงินค่าบัตรเครดิต (จำนวน 17 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_{15} = จำนวนค่าใช้จ่ายในเดือนมิถุนายน ถ้าผลมากกว่า 17386 บาท คือ ไม่ชำระเงินค่าบัตรเครดิต (จำนวน 26 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_4 = สถานภาพการสมรส ถ้าผลแต่งงานแล้ว คือ ชำระเงินค่าบัตรเครดิต (จำนวน 4 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_{22} = จำนวนเงินที่จ่ายในเดือนพฤษภาคม ถ้าผลน้อยกว่าเท่ากับ 1663 บาท คือ ไม่ชำระเงินค่าบัตรเครดิต (จำนวน 7 คน) ผลมากกว่า 1663 บาท คือ ชำระเงินค่าบัตรเครดิต (จำนวน 2 คน)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

22:33:56 - trees.J48

```
Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          63               96.9231 %
Incorrectly Classified Instances         2                3.0769 %
Kappa statistic                          0.9091
Mean absolute error                      0.0568
Root mean squared error                  0.1685
Relative absolute error                 15.8073 %
Root relative squared error             39.9927 %
Total Number of Instances               65

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               1.000    0.133      0.962      1.000    0.980      0.913    0.968     0.982     no
               0.867    0.000      1.000      0.867    0.929      0.913    0.968     0.918     yes
Weighted Avg.  0.969    0.103      0.970      0.969    0.968      0.913    0.968     0.967

=== Confusion Matrix ===

  a  b   <-- classified as
 50  0 |  a = no
  2 13 |  b = yes
```

รูปที่ 4-50 ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 10

จากรูป 4-50 สุ่มครั้งที่ 1 Random Seed = 10 กลุ่มลูกค้า 10 เปอร์เซ็นต์ จำนวนทั้งหมด 65 คน มีค่าทำนายความถูกต้องจำนวน 63 คน คิดเป็น 96.9231% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.1685)^2$ = 0.0284 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0568

รูปที่ 4-51 ต้นไม้เพื่อการตัดสินใจ การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 20

จากรูป 4-51 แสดงต้นไม้เพื่อการตัดสินใจ การชำระเงินด้วยบัตรเครดิตของลูกค้า โหนดรากแบ่งโดยตัวแปรทำนายคือ X_8 = สถานการณ์ชำระเงินในเดือนกรกฎาคม ตัดสินใจต่อด้วยตัวแปรทำนาย X_{18} = จำนวนเงินที่จ่ายในเดือนกันยายน ถ้าผลน้อยกว่าเท่ากับ 3200 บาท คือ ชำระเงินค่า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

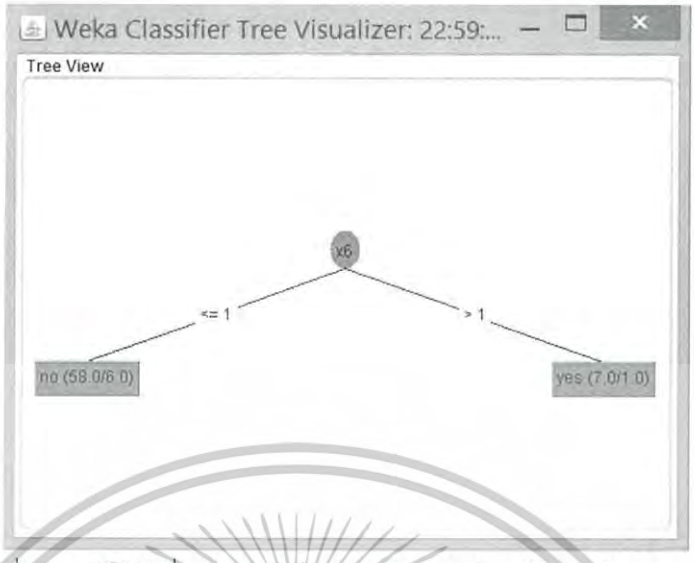
บัตรเครดิต (จำนวน 4 คน) ถ้าผลมากกว่า 3200 บาท คือ ไม่ชำระเงินค่าบัตรเครดิต (จำนวน 2 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_{17} = จำนวนค่าใช้จ่ายในเดือนเมษายน ถ้าผลมากกว่า 20782 บาท คือ ไม่ชำระเงินค่าบัตรเครดิต (จำนวน 30 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_3 = การศึกษา ผลสำเร็จการศึกษาแล้ว คือ ไม่ชำระเงินค่าบัตรเครดิต (จำนวน 9 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_{19} = จำนวนเงินที่จ่ายในเดือนสิงหาคม ถ้าผลมากกว่า 3000 บาท คือ ชำระเงินค่าบัตรเครดิต (จำนวน 5 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_5 = อายุ ถ้าผลมากกว่า 43 ปี คือ ชำระเงินค่าบัตรเครดิต (จำนวน 3 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_{15} = จำนวนค่าใช้จ่ายในเดือนมิถุนายน ถ้าผลน้อยกว่าเท่ากับ 36436 บาท คือ ไม่ชำระค่าบัตรเครดิต (จำนวน 10 คน) ถ้าผลมากกว่า 36436 บาท คือ ชำระเงินค่าบัตรเครดิต (จำนวน 2 คน)

```
00:22:56 - trees.J48

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          63               96.9231 %
Incorrectly Classified Instances         2                3.0769 %
Kappa statistic                          0.9134
Mean absolute error                      0.0571
Root mean squared error                  0.169
Relative absolute error                 15.2262 %
Root relative squared error             39.2165 %
Total Number of Instances               65

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 1.000    0.125      0.961      1.000    0.980      0.917    0.966     0.980     no
                 0.875    0.000      1.000      0.875    0.933      0.917    0.966     0.935     yes
Weighted Avg.    0.969    0.094      0.970      0.969    0.969      0.917    0.966     0.968

=== Confusion Matrix ===

  a  b   <-- classified as
 49  0 |  a = no
  2 14 |  b = yes
```

รูปที่ 4-52 ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 20

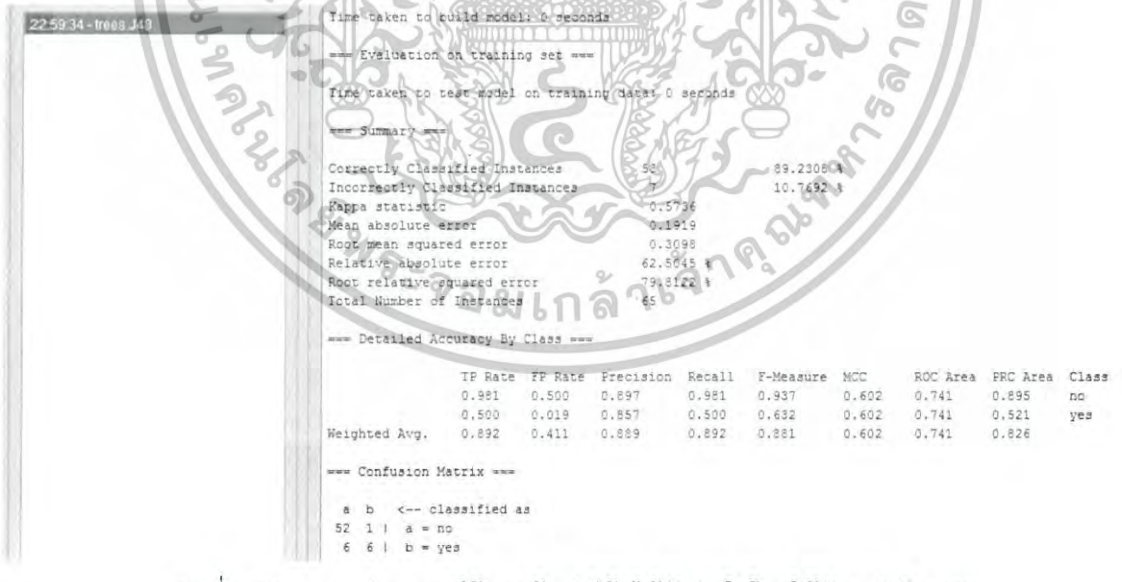
จากรูป 4-52 สุ่มครั้งที่ 2 Random Seed = 20 กลุ่มลูกค้า 10 เปอร์เซ็นต์ จำนวนทั้งหมด 65 คน มีค่าทำนายความถูกต้องจำนวน 63 คน คิดเป็น 96.9231% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.169)^2$ = 0.0286 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0571

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4-53 ต้นไม้เพื่อการตัดสินใจ การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 30

จากรูป 4-53 แสดงต้นไม้เพื่อการตัดสินใจ การชำระเงินด้วยบัตรเครดิตของลูกค้า โหนดราก แบ่งโดยตัวแปรทำนายคือ X_6 = สถานการณ์ชำระเงินในเดือนกันยายน ถ้าผลน้อยกว่าเท่ากับ 1 บาท คือ ไม่ชำระเงินค่าบัตรเครดิต (จำนวน 58 คน) ถ้าผลมากกว่า 1 บาท คือ ชำระเงินค่าบัตรเครดิต (จำนวน 7 คน)



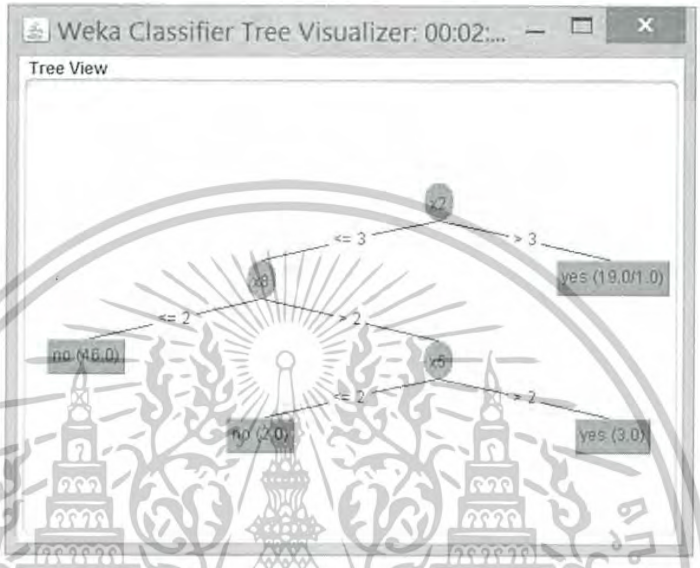
รูปที่ 4-54 ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 10

จากรูป 4-54 สุ่มครั้งที่ 3 Random Seed = 30 กลุ่มลูกค้า 10 เปอร์เซ็นต์ จำนวนทั้งหมด 65 คน มีค่าทำนายความถูกต้องจำนวน 58 คน คิดเป็น 89.2308% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.3098)^2$ = 0.0960 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.1919

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.4.2 ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 โดยโปรแกรม WEKA

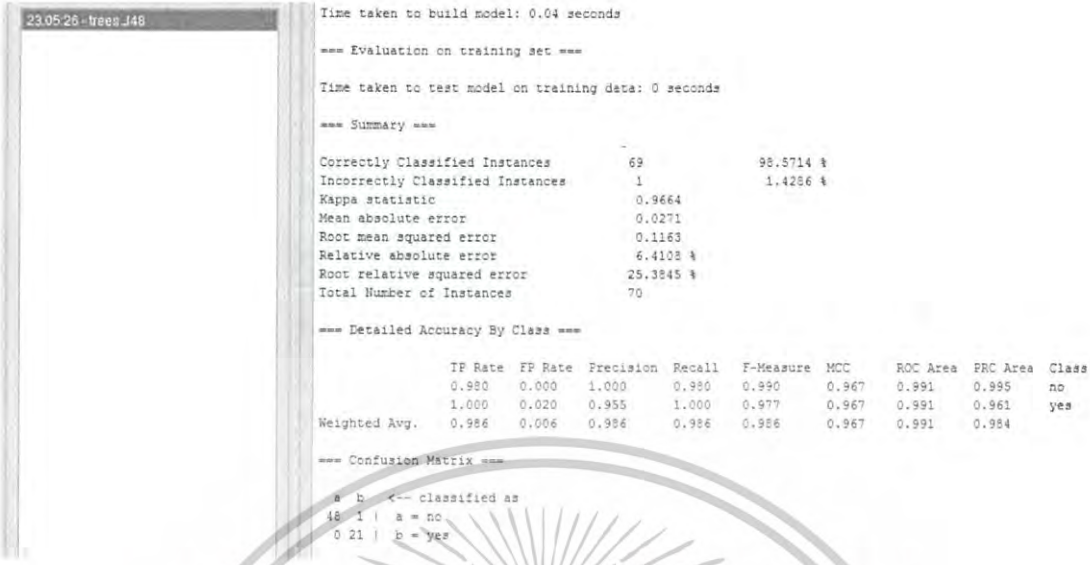
4.4.2.1 โรคมะเร็งเต้านมของรัฐวิสคอนซิน (Breast Cancer Wisconsin) จำนวนข้อมูลทดสอบ 699 ค่า ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่ต่ำ เมื่อกำหนดค่า Random Seed 10 20 และ 30 ตามลำดับ



รูปที่ 4-55 ต้นไม้เพื่อการตัดสินใจ โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 10

จากรูป 4-55 แสดงต้นไม้เพื่อการตัดสินใจ โรคมะเร็งเต้านมของรัฐวิสคอนซิน โหนดรากแบ่งโดยตัวแปรทำนายคือ X_2 = ความสม่ำเสมอของขนาดเซลล์ ถ้าผลมากกว่า 3 นาโนเมตร คือ เป็นโรคมะเร็งเต้านม (จำนวน 19 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_8 = นิวคลีโอไลในภาวะปกติ ถ้าผลน้อยกว่าเท่ากับ 2 เซลล์ คือ ไม่เป็นโรคมะเร็งเต้านม (จำนวน 46 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_5 = ขนาดเซลล์เดียว ถ้าผลน้อยกว่าเท่ากับ 2 มิลลิเมตร คือ ไม่เป็นโรคมะเร็งเต้านม (จำนวน 2 คน) ถ้าผลมากกว่า 2 มิลลิเมตร คือ เป็นโรคมะเร็งเต้านม (จำนวน 3 คน)

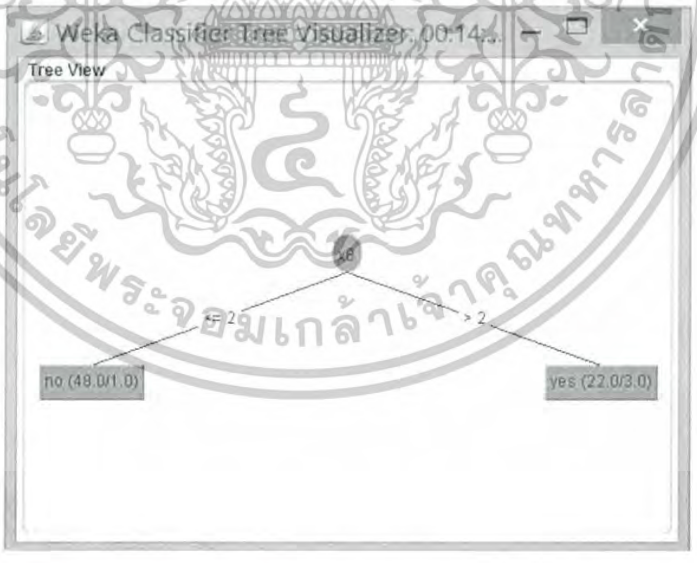
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



```
23.05.26 - trees.J48

Time taken to build model: 0.04 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          69               98.5714 %
Incorrectly Classified Instances         1                1.4286 %
Kappa statistic                          0.9664
Mean absolute error                      0.0271
Root mean squared error                  0.1163
Relative absolute error                  6.4108 %
Root relative squared error             25.3845 %
Total Number of Instances               70

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.980    0.000    1.000      0.980   0.990      0.967    0.991     0.995     no
                 1.000    0.020    0.955      1.000   0.977      0.967    0.991     0.961     yes
Weighted Avg.    0.986    0.006    0.986      0.986   0.986      0.967    0.991     0.984

=== Confusion Matrix ===

  a  b   <-- classified as
 48  1 |  a = no
  0 21 |  b = yes
```

รูปที่ 4-56 ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 10

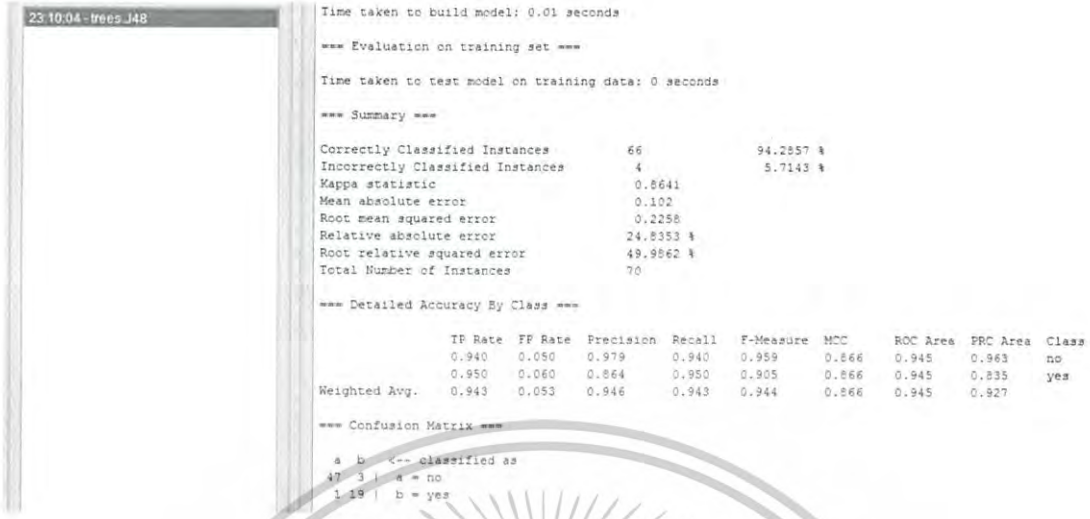
จากรูป 4-56 สุ่มครั้งที่ 1 Random Seed = 10 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 70 คน มีค่าทำนายความถูกต้องจำนวน 69 คน คิดเป็น 98.5714% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.1163)^2$ = 0.0135 และมีค่าสัมบูรณ์เฉลี่ย = 0.0271



รูปที่ 4-57 ต้นไม้เพื่อการตัดสินใจ โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 20

จากรูป 4-57 แสดงต้นไม้เพื่อการตัดสินใจ โรคมะเร็งเต้านมของรัฐวิสคอนซิน โหนดรากแบ่งโดยตัวแปรทำนายคือ X_8 = นิวคลีโอไลในภาวะปกติ ถ้าผลน้อยกว่าเท่ากับ 2 เซลล์ คือ ไม่เป็นโรคมะเร็งเต้านม (จำนวน 48 คน) ถ้าผลมากกว่า 2 เซลล์ คือ เป็นโรคมะเร็งเต้านม (จำนวน 22 คน)

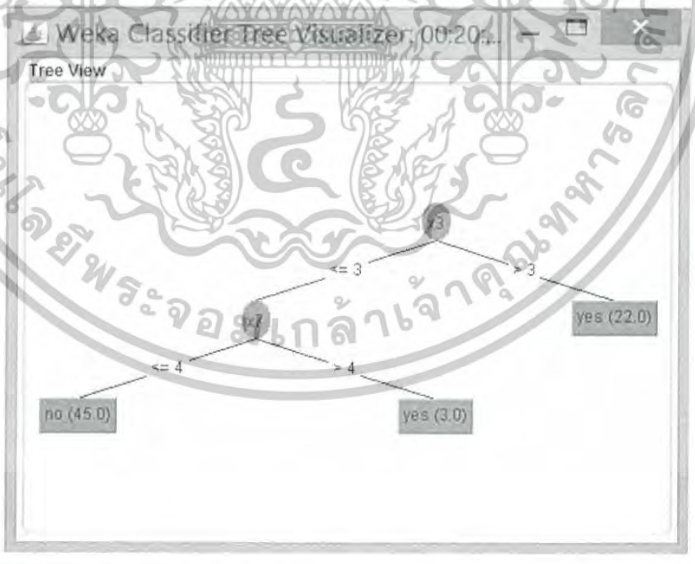
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



```
23:10:04 - trees.J48

Time taken to build model: 0.01 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          66               94.2857 %
Incorrectly Classified Instances         4                5.7143 %
Kappa statistic                          0.8641
Mean absolute error                      0.102
Root mean squared error                  0.2258
Relative absolute error                 24.8353 %
Root relative squared error             49.9862 %
Total Number of Instances               70

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.940    0.050    0.979      0.940   0.959      0.866    0.945     0.963     no
                 0.950    0.060    0.864      0.950   0.905      0.866    0.945     0.835     yes
Weighted Avg.    0.943    0.053    0.946      0.943   0.944      0.866    0.945     0.927

=== Confusion Matrix ===

  a  b   <-- classified as
 47  3 |  a = no
  1 19 |  b = yes
```

รูปที่ 4-58 ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 20

จากรูป 4-58 สุ่มครั้งที่ 2 Random Seed = 20 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 70 คน มีค่าทำนายความถูกต้องจำนวน 66 คน คิดเป็น 94.2857% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.2258)^2$ = 0.0501 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.102

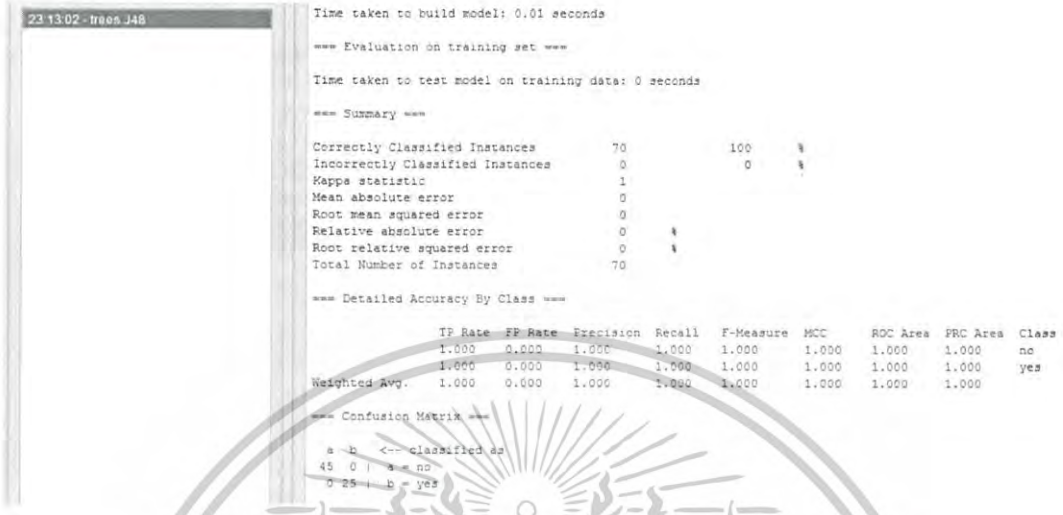


รูปที่ 4-59 ต้นไม้เพื่อการตัดสินใจ โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 30

จากรูป 4-59 แสดงต้นไม้เพื่อการตัดสินใจ โรคมะเร็งเต้านมของรัฐวิสคอนซิน โหนดรากแบ่งโดยตัวแปรทำนายคือ X_3 = ความสม่ำเสมอของรูปร่างเซลล์ ถ้าผลมากกว่า 3 ยูนิตเซลล์ คือ เป็นโรคมะเร็งเต้านม (จำนวน 22 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_7 = ของเหลวในเซลล์ ถ้าผลน้อย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

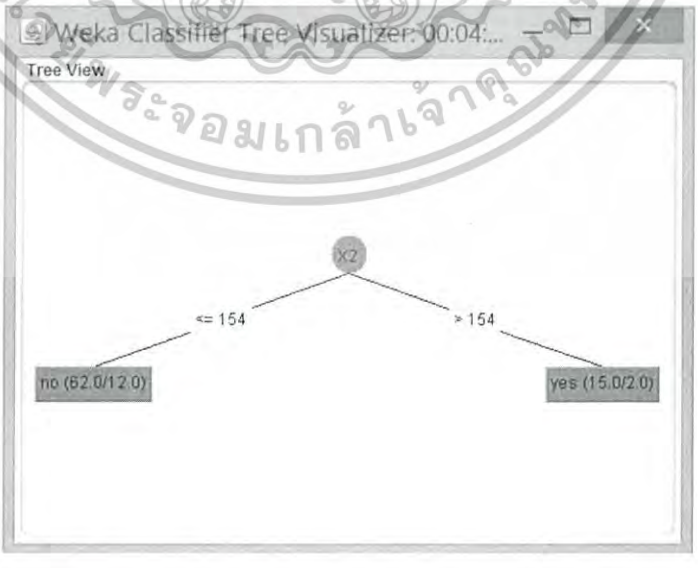
กว่าเท่ากับ 4 นาโนเมตร คือ ไม่เป็นโรคมะเร็งเต้านม (จำนวน 45 คน) ถ้าผลมากกว่า 4 นาโนเมตร คือ เป็นโรคมะเร็งเต้านม (จำนวน 3 คน)



รูปที่ 4-60 ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 30

จากรูป 4-60 สุ่มครั้งที่ 3 Random Seed = 30 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 70 คน มีค่าทำนายความถูกต้องจำนวน 70 คน คิดเป็น 100% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0)^2$ = 0 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0

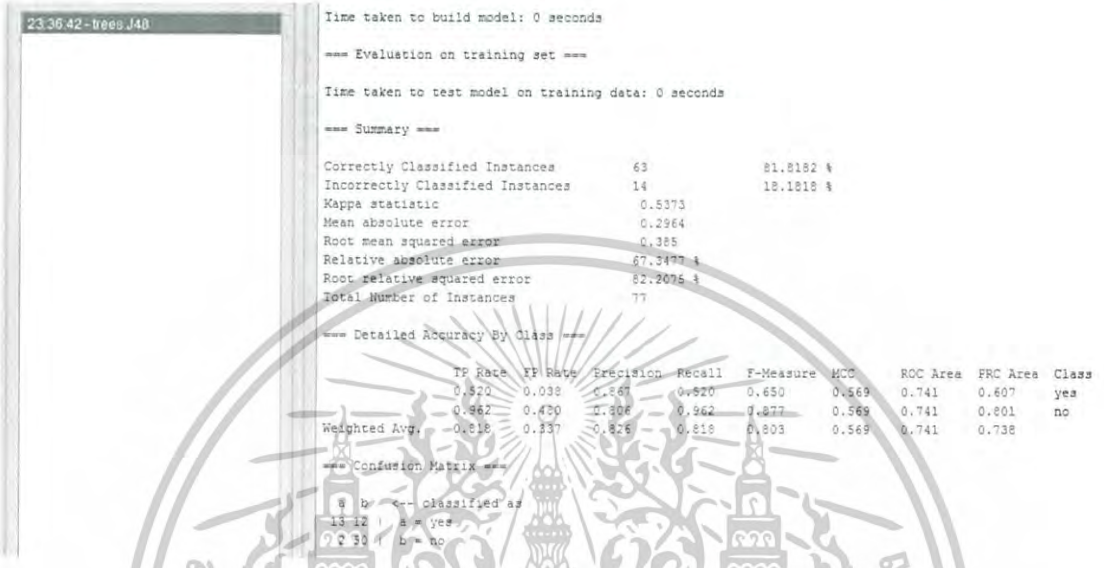
4.4.2.2 โรคเบาหวานของชาวพิม่า ประเทศอินเดีย (Pima Indians Diabetes) จำนวนข้อมูลทดสอบ 768 ค่า ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่ปานกลาง เมื่อกำหนดค่า Random Seed 10 20 และ 30 ตามลำดับ



รูปที่ 4-61 ต้นไม้เพื่อการตัดสินใจ โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 10

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

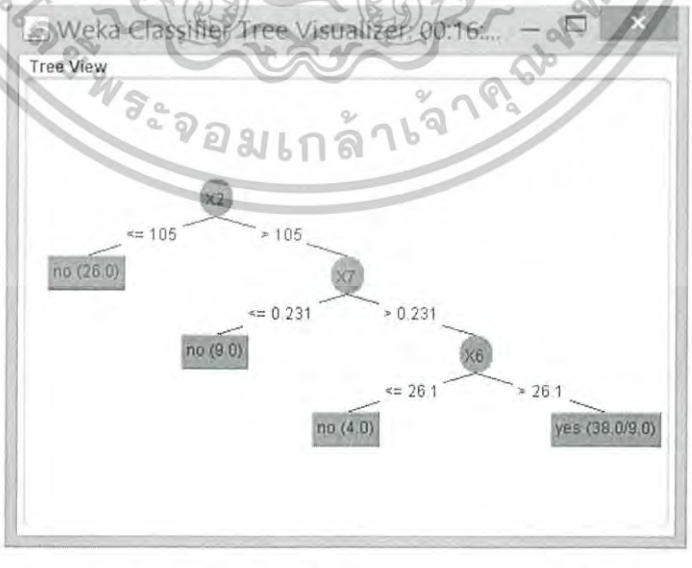
จากรูป 4-75 แสดงต้นไม้เพื่อการตัดสินใจ โรคเบาหวานของชาวพิม่า ประเทศอินเดีย โหนดรากแบ่งโดยตัวแปรทำนายคือ X_2 = ความเข้มข้นของกลูโคสในเลือด 2 ชั่วโมงก่อนการทดสอบระดับกลูโคสในช่องปาก ถ้าผลน้อยกว่าเท่ากับ 154 มิลลิกรัมต่อเดซิเบล คือ ไม่เป็นโรคเบาหวาน (จำนวน 62 คน) ถ้าผลมากกว่า 154 มิลลิกรัมต่อเดซิเบล คือ เป็นโรคเบาหวาน (จำนวน 15 คน)



```
23:36:42 - trees.J48

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          63               81.8182 %
Incorrectly Classified Instances        14               18.1818 %
Kappa statistic                          0.5373
Mean absolute error                      0.2964
Root mean squared error                  0.385
Relative absolute error                 67.3477 %
Root relative squared error             82.2076 %
Total Number of Instances               77

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0.520     0.038     0.867     0.520     0.650     0.569     0.741     0.607     yes
                 0.962     0.480     0.806     0.962     0.877     0.569     0.741     0.801     no
Weighted Avg.    0.818     0.337     0.826     0.818     0.803     0.569     0.741     0.738

=== Confusion Matrix ===

  a  b   <-- classified as
 13 12 |  a = yes
  2 50 |  b = no
```

รูปที่ 4-62 ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 10

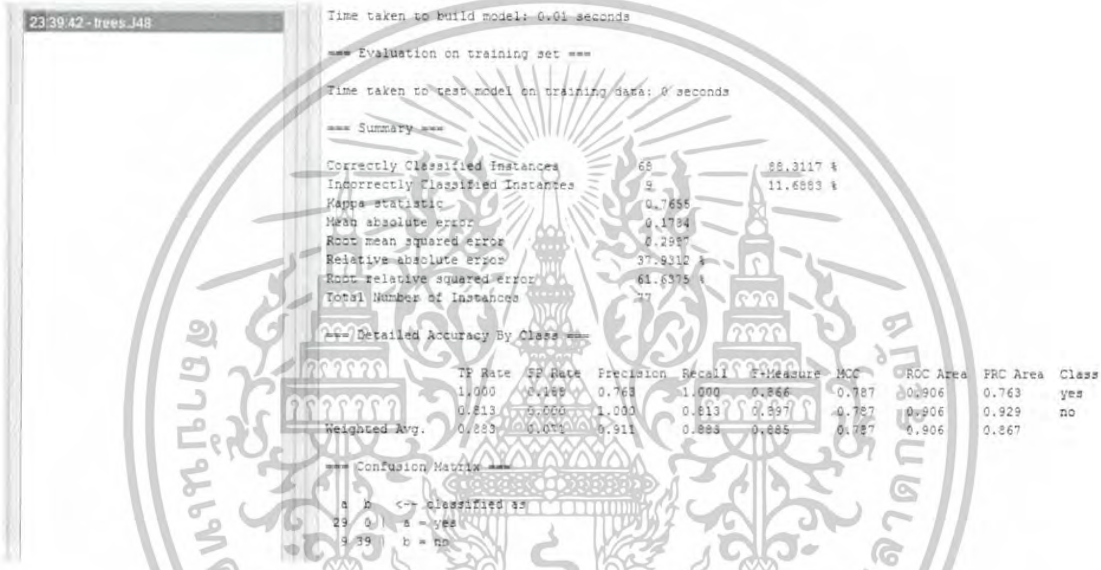
จากรูป 4-62 สุ่มครั้งที่ 1 Random Seed = 10 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 77 คน มีค่าทำนายความถูกต้องจำนวน 63 คน คิดเป็น 81.8182% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.385)^2$ = 0.1482 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.2964



รูปที่ 4-63 ต้นไม้เพื่อการตัดสินใจ โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 20

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากรูป 4-63 แสดงต้นไม้เพื่อการตัดสินใจ โรคเบาหวานของชาวพิม่า ประเทศอินเดีย โหนดรากแบ่งโดยตัวแปรทำนายคือ X_2 = ความเข้มข้นของกลูโคสในเลือด 2 ชั่วโมงก่อนการทดสอบระดับกลูโคสในช่องปาก ถ้าผลน้อยกว่าเท่ากับ 105 มิลลิกรัมต่อเดซิเบล คือ ไม่เป็นโรคเบาหวาน (จำนวน 26 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_7 = พันธุกรรมของผู้ป่วยที่มีความเสี่ยงเป็นโรคเบาหวาน ถ้าผลน้อยกว่าเท่ากับร้อยละ 0.231 คือ ไม่เป็นโรคเบาหวาน (จำนวน 9 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_6 = ค่าดัชนีมวลกาย ถ้าผลน้อยกว่าเท่ากับ 26.1 กิโลกรัมต่อตารางเมตร คือ ไม่เป็นโรคเบาหวาน (จำนวน 4 คน) ถ้าผลมากกว่า 26.1 กิโลกรัมต่อตารางเมตร คือ เป็นโรคเบาหวาน (จำนวน 38 คน)



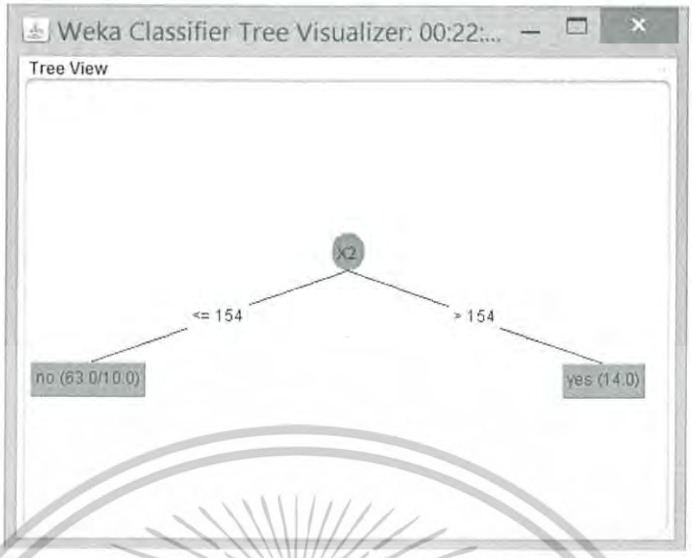
23:39:42 - trees.J48

```
Time taken to build model: 0.01 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          68               88.3117 %
Incorrectly Classified Instances         9               11.6883 %
Kappa statistic                          0.7655
Mean absolute error                      0.1784
Root mean squared error                  0.2987
Relative absolute error                 37.9312 %
Root relative squared error             61.6375 %
Total Number of Instances               77

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 1.000    0.188      0.763     1.000     0.866      0.787    0.906     0.763     yes
                 0.813    0.000      1.000     0.813     0.897      0.787    0.906     0.929     no
Weighted Avg.    0.883    0.071      0.911     0.883     0.885      0.787    0.906     0.867

=== Confusion Matrix ===

  a  b   <-- classified as
 29  0 |  a = yes
  9 39 |  b = no
```

รูปที่ 4-64 ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 20

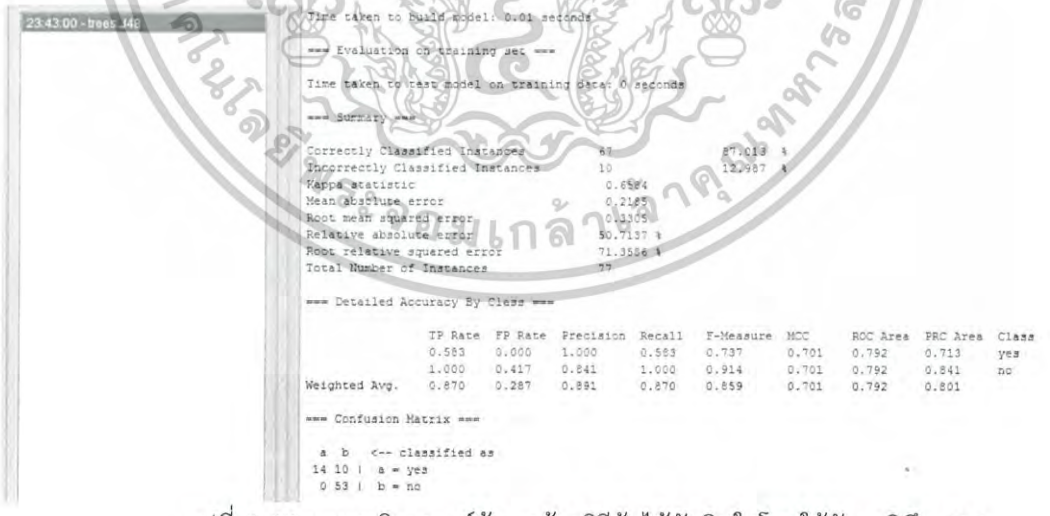
จากรูป 4-64 สุ่มครั้งที่ 2 Random Seed = 20 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 77 คน มีค่าทำนายความถูกต้องจำนวน 68 คน คิดเป็น 88.3117% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.2987)^2$ = 0.0892 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.1784

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4-65 ต้นไม้เพื่อการตัดสินใจ โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 30

จากรูป 4-65 แสดงต้นไม้เพื่อการตัดสินใจ โรคเบาหวานของชาวพิม่า ประเทศอินเดีย โหนดรากแบ่งโดยตัวแปรทำนายคือ X_2 = ความเข้มข้นของกลูโคสในเลือด 2 ชั่วโมงก่อนการทดสอบระดับกลูโคสในช่องปาก ถ้าผลน้อยกว่าเท่ากับ 154 มิลลิกรัมต่อเดซิเบล คือ ไม่เป็นโรคเบาหวาน (จำนวน 63 คน) ถ้าผลมากกว่า 154 มิลลิกรัมต่อเดซิเบล คือ เป็นโรคเบาหวาน (จำนวน 14 คน)

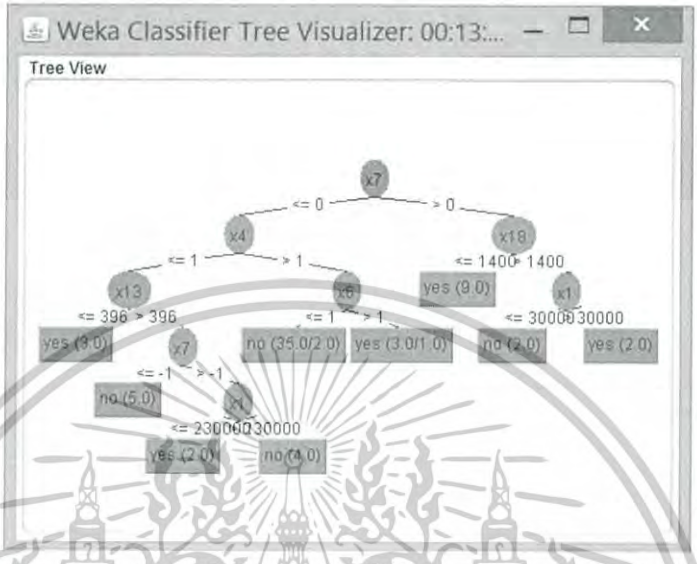


รูปที่ 4-66 ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 30

จากรูป 4-66 สุ่มครั้งที่ 3 Random Seed = 30 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 77 คน มีค่าทำนายความถูกต้องจำนวน 67 คน คิดเป็น 87.013% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.3305)^2$ = 0.1092 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.2185

เอกสารนี้ใช้เพื่อการศึกษาเท่านั้น ไม่อนุญาตให้ทำซ้ำ ดัดแปลง เผยแพร่ หรือใช้ประโยชน์ทางการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ลบเลือนหรือดัดแปลงข้อความนี้ในเอกสารทุกครั้งที่มีการนำไปใช้

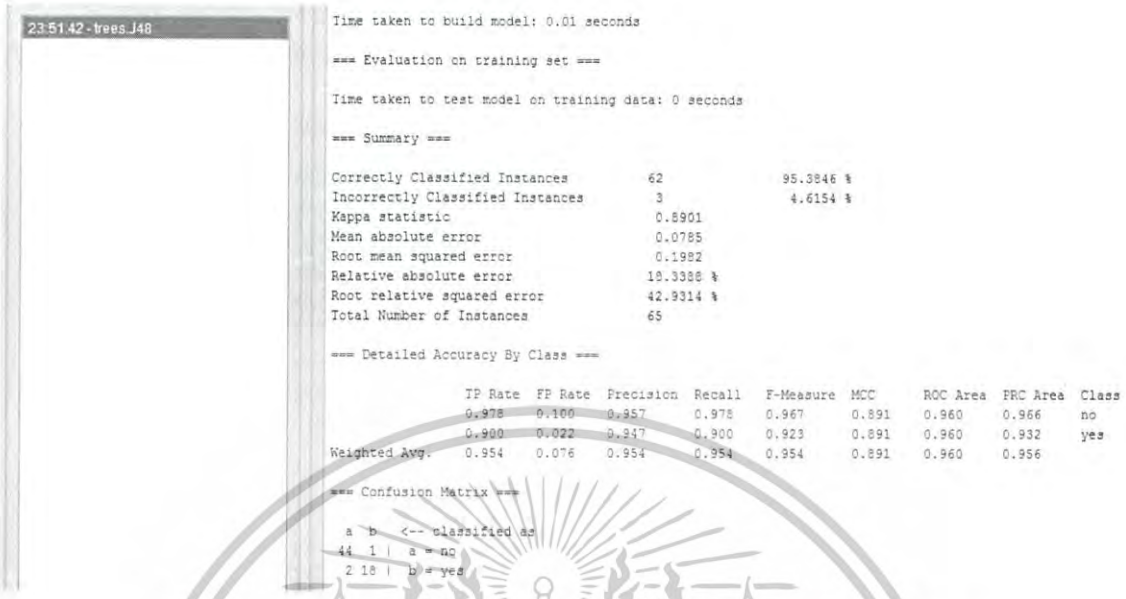
4.4.2.3 การชำระเงินด้วยบัตรเครดิตของลูกค้า (Default of Credit Card Clients) จำนวนข้อมูลทดสอบ 646 ค่า ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่สูง เมื่อกำหนดค่า Random Seed 10 20 และ 30 ตามลำดับ



รูปที่ 4-67 ต้นไม้เพื่อการตัดสินใจ การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 10

จากรูป 4-67 แสดงต้นไม้เพื่อการตัดสินใจ การชำระเงินด้วยบัตรเครดิตของลูกค้า โหนดรากแบ่งโดยตัวแปรทำนายคือ X_7 = สถานการณ์ชำระเงินในเดือนสิงหาคม ตัดสินใจต่อด้วยตัวแปรทำนาย X_{13} = จำนวนค่าใช้จ่ายในเดือนสิงหาคม ถ้าผลน้อยกว่าเท่ากับ 396 บาท คือ ชำระเงินค่าบัตรเครดิต (จำนวน 3 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_7 = สถานการณ์ชำระเงินในเดือนสิงหาคม ผลชำระเงินตรงเวลาหรือล่วงหน้า คือ ไม่ชำระเงินค่าบัตรเครดิต (จำนวน 5 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_1 = จำนวนเครดิตที่ได้รับ รวมทั้งสินเชื่อผู้บริโภคแต่ละรายและเครดิตครอบครัว ถ้าผลน้อยกว่าเท่ากับ 230000 TWD คือ ชำระเงินค่าบัตรเครดิต (จำนวน 2 คน) ถ้าผลมากกว่า 230000 TWD คือ ไม่ชำระค่าบัตรเครดิต (จำนวน 4 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_6 = สถานการณ์ชำระเงินในเดือนกันยายน ถ้าผลชำระล่าช้า 1 เดือน คือ ไม่ชำระเงินค่าบัตรเครดิต (จำนวน 35 คน) ถ้าผลชำระเงินล่าช้ามากกว่า 1 เดือน คือ ชำระเงินค่าบัตรเครดิต (จำนวน 3 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_{18} = จำนวนเงินที่จ่ายในเดือนกันยายน ผลน้อยกว่าเท่ากับ1400 บาท คือ ชำระเงินค่าบัตรเครดิต (จำนวน 9 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_1 = จำนวนเครดิตที่ได้รับ รวมทั้งสินเชื่อผู้บริโภคแต่ละรายและเครดิตครอบครัว ถ้าผลน้อยกว่าเท่ากับ 30000 TWD คือ ชำระเงินค่าบัตรเครดิต (จำนวน 2 คน) ถ้าผลมากกว่า 30000 TWD คือ ไม่ชำระค่าบัตรเครดิต (จำนวน 2 คน)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

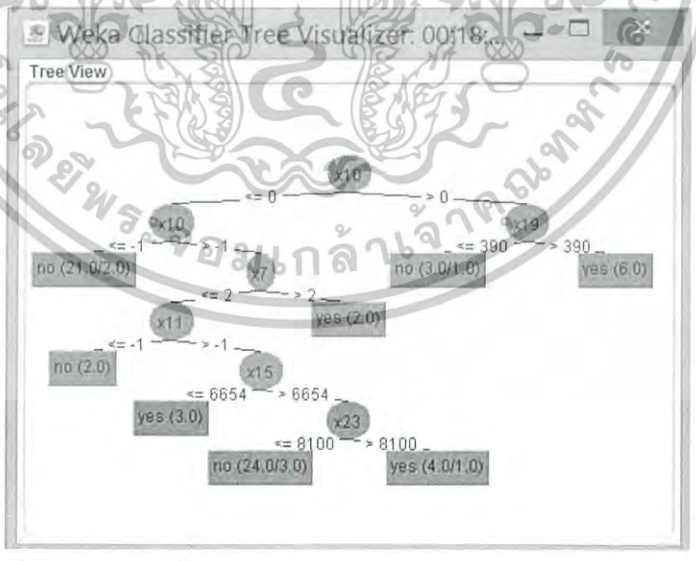


23:51:42 - trees.J48

```
Time taken to build model: 0.01 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          62               95.3846 %
Incorrectly Classified Instances         3                4.6154 %
Kappa statistic                          0.8901
Mean absolute error                      0.0785
Root mean squared error                  0.1982
Relative absolute error                 18.3388 %
Root relative squared error             42.9314 %
Total Number of Instances               65

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.978    0.100    0.957      0.978   0.967      0.891    0.960     0.966     no
               0.900    0.022    0.947      0.900   0.923      0.891    0.960     0.932     yes
Weighted Avg.  0.954    0.076    0.954      0.954   0.954      0.891    0.960     0.956

=== Confusion Matrix ===

  a  b   <-- classified as
 44  1 |  a = no
  2 18 |  b = yes
```

รูปที่ 4-68 ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 10

จากรูป 4-68 สุ่มครั้งที่ 1 Random Seed = 10 กลุ่มลูกค้า 10 เปอร์เซ็นต์ จำนวนทั้งหมด 65 คน มีค่าทำนายความถูกต้องจำนวน 62 คน คิดเป็น 95.3846% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.1982)^2$ = 0.0393 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0785



รูปที่ 4-69 ต้นไม้เพื่อการตัดสินใจ การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 20

จากรูป 4-69 แสดงต้นไม้เพื่อการตัดสินใจ การชำระเงินด้วยบัตรเครดิตของลูกค้า โหนดราก แบ่งโดยตัวแปรทำนายคือ X_{10} = สถานการณ์ชำระเงินในเดือนพฤษภาคม ถ้าผลชำระเงินตรงเวลา

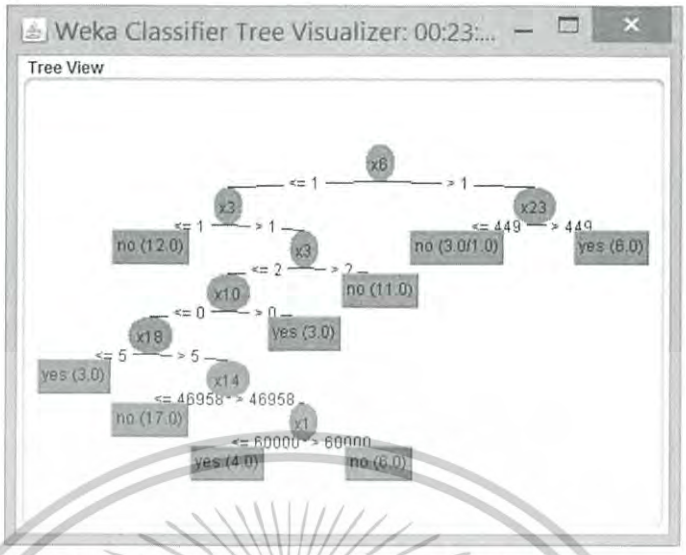
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

หรือล่วงหน้า คือ ไม่ชำระเงินค่าบัตรเครดิต (จำนวน 21 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_7 = สถานการณ์ชำระเงินในเดือนสิงหาคม ผลชำระเงินล่าช้ามากกว่า2 เดือน คือ ชำระเงินค่าบัตรเครดิต (จำนวน 2 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_{11} = สถานการณ์ชำระเงินในเดือนเมษายน ถ้าผลชำระเงินตรงเวลาหรือล่วงหน้า คือ ไม่ชำระเงินค่าบัตรเครดิต (จำนวน 2 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_{15} = จำนวนค่าใช้จ่ายในเดือนมิถุนายน ผลน้อยกว่าเท่ากับ 6654 บาท คือ ชำระเงินค่าบัตรเครดิต (จำนวน 3 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_{23} = จำนวนเงินที่จ่ายในเดือนเมษายน ถ้าผลน้อยกว่าเท่ากับ 8100 บาท คือ ไม่ชำระเงินค่าบัตรเครดิต (จำนวน 24 คน) ถ้าผลมากกว่า 8100 บาท คือ ชำระเงินค่าบัตรเครดิต (จำนวน 4 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_{19} = จำนวนเงินที่จ่ายในเดือนสิงหาคม ถ้าผลน้อยกว่าเท่ากับ 390 บาท คือ ไม่ชำระเงินค่าบัตรเครดิต (จำนวน 3 คน) ผลมากกว่า 390 คือ ชำระเงินค่าบัตรเครดิต (จำนวน 6 คน)

รูปที่ 4-70 ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 20

จากรูป 4-70 สุ่มครั้งที่ 2 Random Seed = 20 กลุ่มลูกค้า 10 เปอร์เซ็นต์ จำนวนทั้งหมด 65 คน มีค่าทำนายความถูกต้องจำนวน 58 คน คิดเป็น 89.2308% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.3)^2$ = 0.0900 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.18

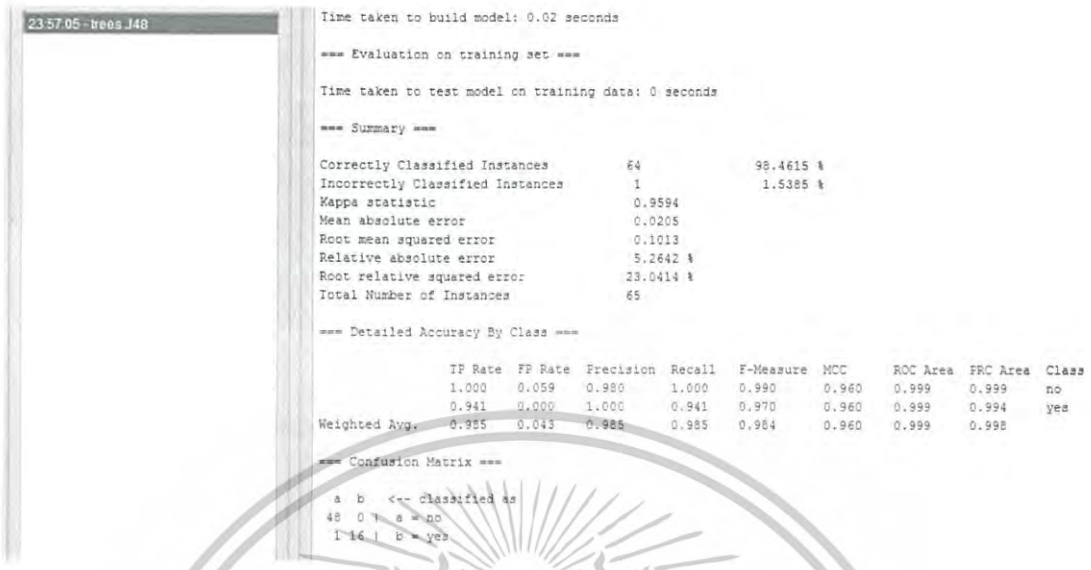
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4-71 ต้นไม้เพื่อการตัดสินใจ การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 30

จากรูป 4-71 แสดงต้นไม้เพื่อการตัดสินใจ การชำระเงินด้วยบัตรเครดิตของลูกค้า โหนดรากแบ่งโดยตัวแปรทำนายคือ X_6 = สถานการณ์ชำระเงินในเดือนกันยายน ตัดสินใจต่อด้วยตัวแปรทำนาย X_3 = การศึกษา ถ้าผลสำเร็จการศึกษา คือ ไม่ชำระเงินค่าบัตรเครดิต (จำนวน 12 คน) ผลยังไม่สำเร็จการศึกษา คือ ไม่ชำระเงินค่าบัตรเครดิต (จำนวน 11 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_{10} = สถานการณ์ชำระเงินในเดือนพฤษภาคม ถ้าผลชำระเงินตรงเวลาหรือล่าช้า คือ ชำระเงินค่าบัตรเครดิต (จำนวน 3 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_{18} = จำนวนเงินที่จ่ายในเดือนกันยายน ผลน้อยกว่าเท่ากับ 5 บาท คือ ชำระเงินค่าบัตรเครดิต (จำนวน 3 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_{14} = จำนวนค่าใช้จ่ายในเดือนกรกฎาคม ผลน้อยกว่าเท่ากับ 46958 บาท คือ ไม่ชำระเงินค่าบัตรเครดิต (จำนวน 17 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_1 = จำนวนเครดิตที่ได้รับ รวมทั้งสินเชื่อผู้บริโภคแต่ละรายและเครดิตครอบครัว ถ้าผลน้อยกว่าเท่ากับ 60000 TWD คือ ชำระเงินค่าบัตรเครดิต (จำนวน 4 คน) ถ้าผลมากกว่า 60000 TWD คือ ไม่ชำระค่าบัตรเครดิต (จำนวน 6 คน) ตัดสินใจต่อด้วยตัวแปรทำนาย X_{23} = จำนวนเงินที่จ่ายในเดือนเมษายน ถ้าผลน้อยกว่าเท่ากับ 449 บาท คือ ไม่ชำระเงินค่าบัตรเครดิต (จำนวน 3 คน) ถ้าผลมากกว่า 449 บาท คือ ชำระเงินค่าบัตรเครดิต (จำนวน 6 คน)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



23:57:05 - trees.J48

Time taken to build model: 0.02 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances	64	98.4615 %
Incorrectly Classified Instances	1	1.5385 %
Kappa statistic	0.9594	
Mean absolute error	0.0205	
Root mean squared error	0.1013	
Relative absolute error	5.2642 %	
Root relative squared error	23.0414 %	
Total Number of Instances	65	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.059	0.980	1.000	0.990	0.960	0.999	0.999	no
	0.941	0.000	1.000	0.941	0.970	0.960	0.999	0.994	yes
Weighted Avg.	0.985	0.043	0.985	0.985	0.984	0.960	0.999	0.998	

=== Confusion Matrix ===

a	b	<-- classified as
48	0	a = no
1	16	b = yes

รูปที่ 4-72 ผลการวิเคราะห์ข้อมูลด้วยวิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึม J48 การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 30

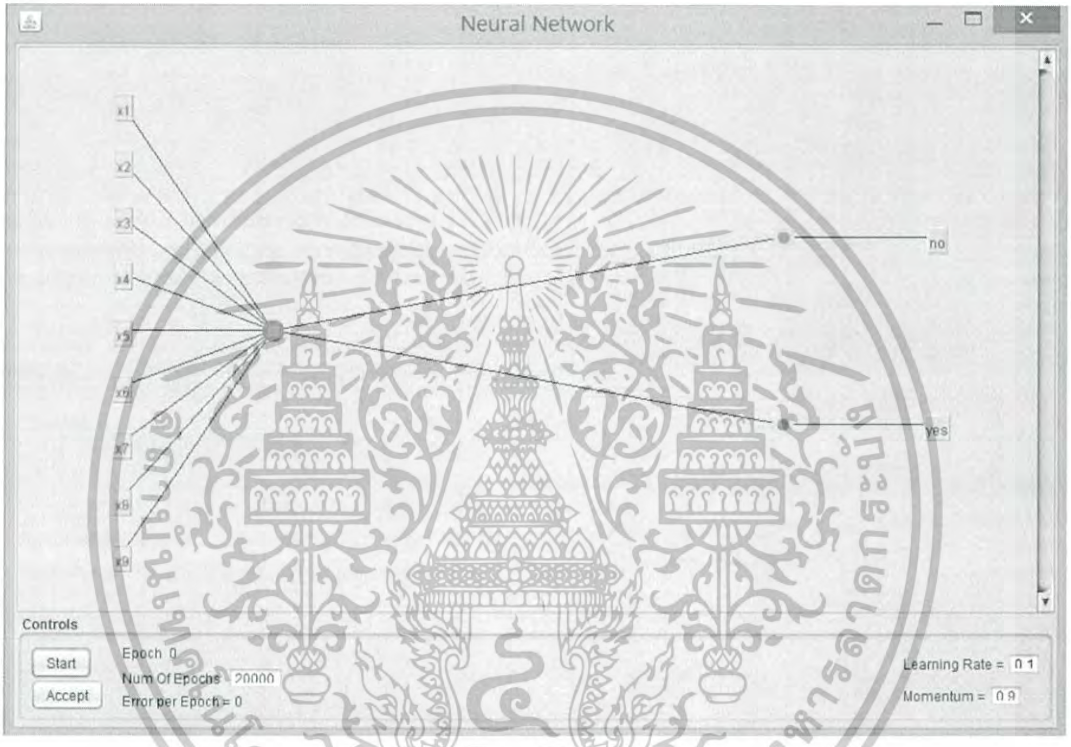
จากรูป 4-72 สุ่มครั้งที่ 3 Random Seed = 30 กลุ่มลูกค้า 10 เปอร์เซ็นต์ จำนวนทั้งหมด 65 คน มีค่าทำนายความถูกต้องจำนวน 64 คน คิดเป็น 98.4615% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.1013)^2$ = 0.0103 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0205

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.5 ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียม โดยใช้อัลกอริทึม Multilayer Perceptron

4.5.1 ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron โดยโปรแกรม SPSS

4.5.1.1 โรคมะเร็งเต้านมของรัฐวิสคอนซิน (Breast Cancer Wisconsin) จำนวนข้อมูลทดสอบ 699 ค่า ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่ต่ำ เมื่อกำหนดค่า Random Seed 10 20 และ 30 ตามลำดับ



รูปที่ 4-73 โครงข่ายประสาท โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 10

จากรูป 4-73 โครงข่ายประสาท โรคมะเร็งเต้านมของรัฐวิสคอนซิน ชั้นข้อมูลเข้า อยู่ด้านซ้ายสุดซึ่งมีคุณลักษณะทั้งหมด 9 ตัวแปร (X_1-X_9) ชั้นซ่อน มี 1 ชั้นซ่อน ชั้นข้อมูลออกหรือชั้นข้อมูลผลลัพธ์ อยู่ด้านขวามือมีคำตอบคือ Yes กับ No Epoch จะรันการทำงานตั้งแต่รอบที่ 1 ถึง 20,000 Error per Epoch แสดงความคลาดเคลื่อนต่อรอบ คือ Learning Rate เท่ากับ 0.1 และ Momentum เท่ากับ 0.9

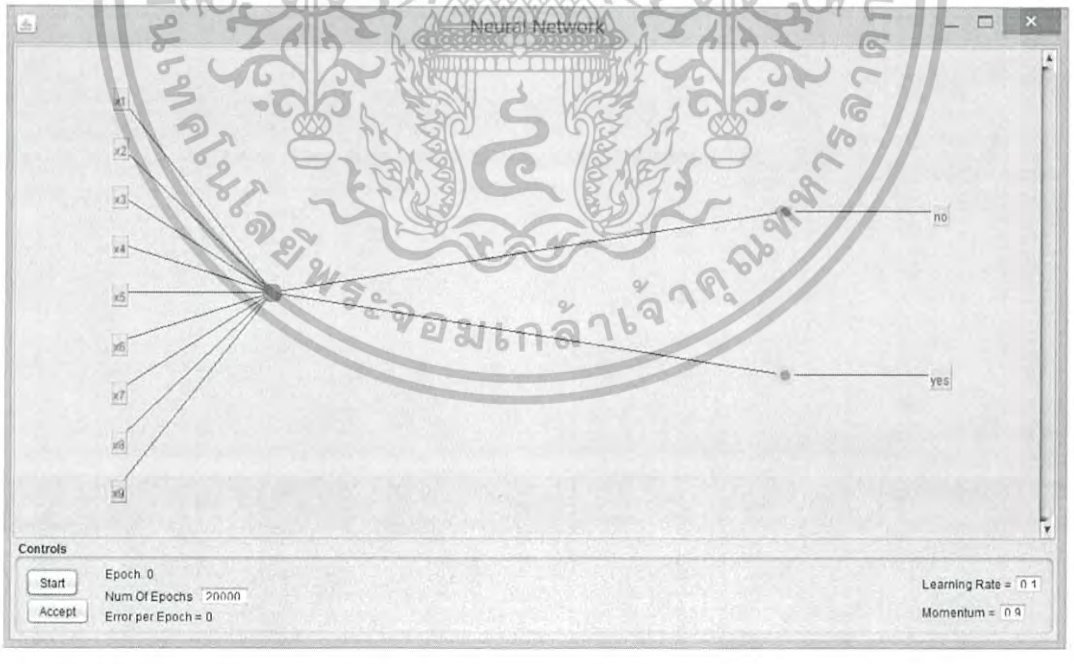
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

23:50:52 - functions.MultilayerPerceptron

```
Time taken to build model: 0.94 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          70               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0013
Root mean squared error                  0.0014
Relative absolute error                  0.2858 %
Root relative squared error              0.2875 %
Total Number of Instances               70

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     no
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     yes
Weighted Avg.  1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

  a  b   <-- classified as
 47  0 |  a = no
  0 23 |  b = yes
```

รูปที่ 4-74 ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 10

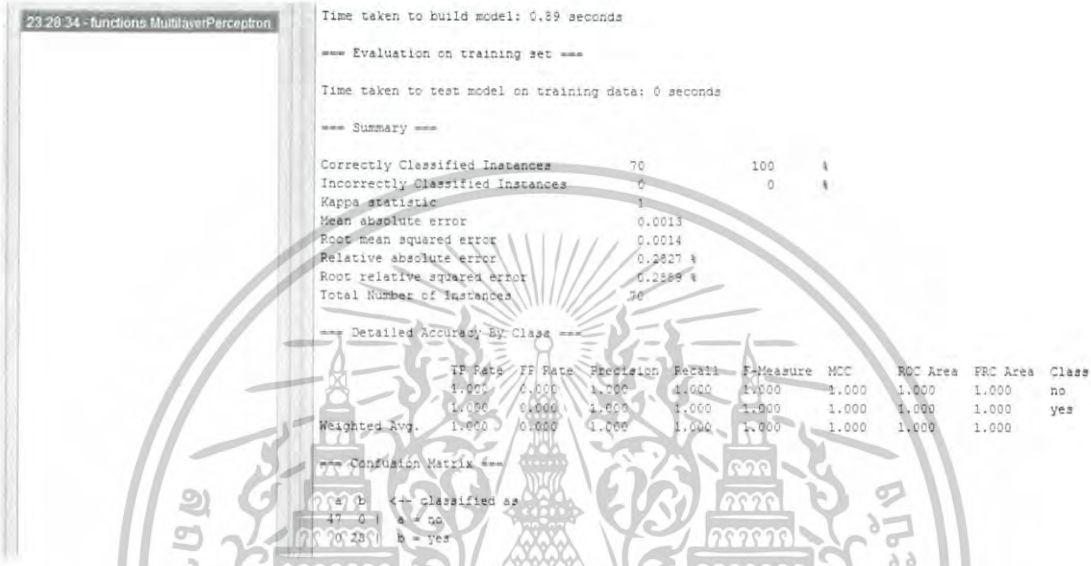
จากรูป 4-74 สุ่มครั้งที่ 1 Random Seed = 10 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 70 คน มีค่าทำนายความถูกต้องจำนวน 70 คน คิดเป็น 100% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.0014)^2$ = 0.0000 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0013



รูปที่ 4-75 โครงข่ายประสาท โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 20

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

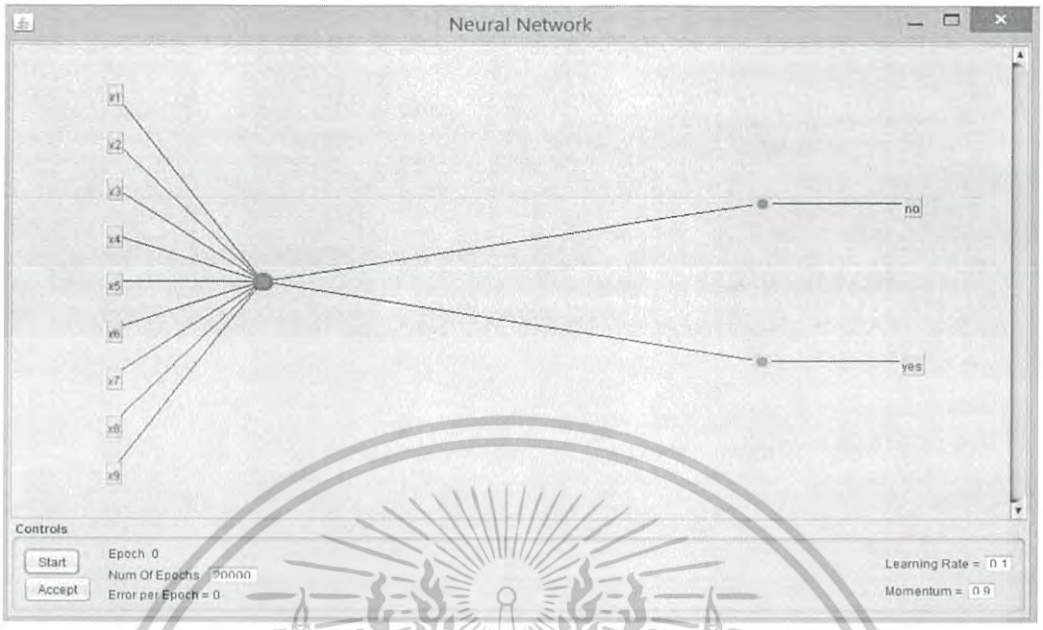
จากรูป 4-75 โครงข่ายประสาท โรคมะเร็งเต้านมของรัฐวิสคอนซิน ชั้นข้อมูลเข้า อยู่ด้านซ้ายสุดซึ่งมีคุณลักษณะทั้งหมด 9 ตัวแปร (X_1-X_9) ชั้นซ่อน มี 1 ชั้นซ่อน ชั้นข้อมูลออกหรือชั้นข้อมูลผลลัพธ์ อยู่ด้านขวามือมีคำตอบคือ Yes กับ No Epoch จะรันการทำงานตั้งแต่รอบที่ 1 ถึง 20,000 Error per Epoch แสดงความคลาดเคลื่อนต่อรอบ คือ Learning Rate เท่ากับ 0.1 และ Momentum เท่ากับ 0.9



รูปที่ 4-76 ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 20

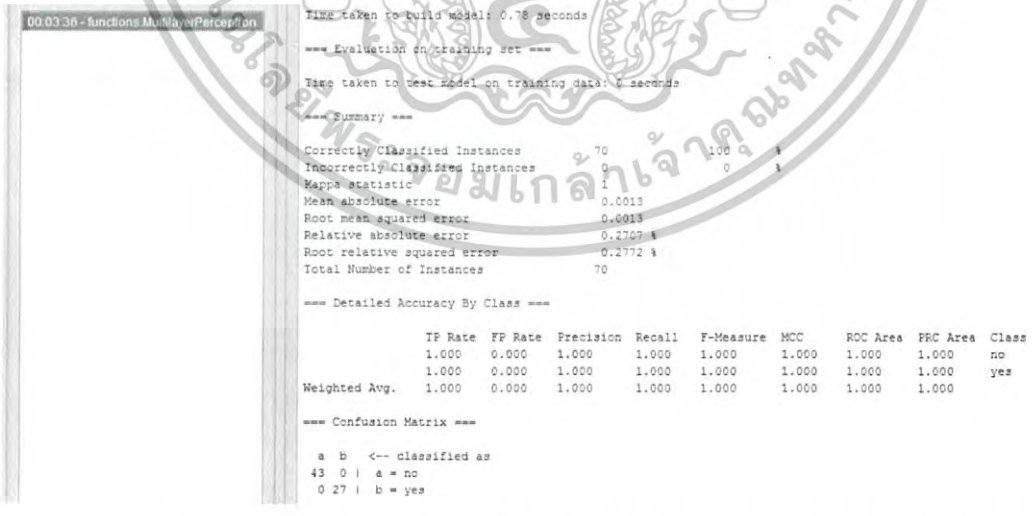
จากรูป 4-76 สุ่มครั้งที่ 2 Random Seed = 20 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 70 คน มีค่าทำนายความถูกต้องจำนวน 70 คน คิดเป็น 100% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.0014)^2$ = 0.0000 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0013

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4-77 โครงข่ายประสาท โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 30

จากรูป 4-77 โครงข่ายประสาท โรคมะเร็งเต้านมของรัฐวิสคอนซิน ชั้นข้อมูลเข้า อยู่ด้านซ้ายสุดซึ่งมีคุณลักษณะทั้งหมด 9 ตัวแปร (X_1-X_9) ชั้นซ่อน มี 1 ชั้นซ่อน ชั้นข้อมูลออกหรือชั้นข้อมูลผลลัพธ์ อยู่ด้านขวามือมีคำตอบคือ Yes กับ No Epoch จะรันการทำงานตั้งแต่รอบที่ 1 ถึง 20,000 Error per Epoch แสดงความคลาดเคลื่อนต่อรอบ คือ Learning Rate เท่ากับ 0.1 และ Momentum เท่ากับ 0.9



00:03:36 - functions.MultilayerPerceptron

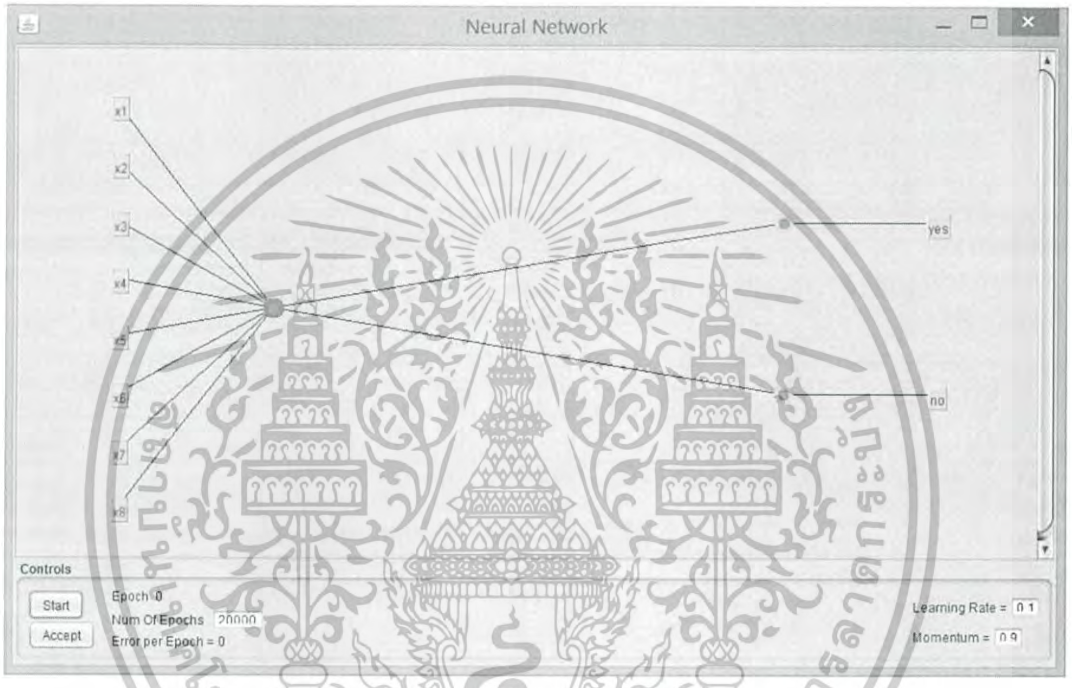
```
Time taken to build model: 0.78 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          70               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0013
Root mean squared error                  0.0013
Relative absolute error                  0.2707 %
Root relative squared error              0.2772 %
Total Number of Instances               70

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     no
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     yes
Weighted Avg.  1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

  a  b   <-- classified as
 43  0 |  a = no
  0 27 |  b = yes
```

รูปที่ 4-78 ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 30

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากรูป 4-78 สุ่มครั้งที่ 3 Random Seed = 30 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 70 คน มีค่าทำนายความถูกต้องจำนวน 70 คน คิดเป็น 100% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.0013)^2$ = 0.0000 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0013

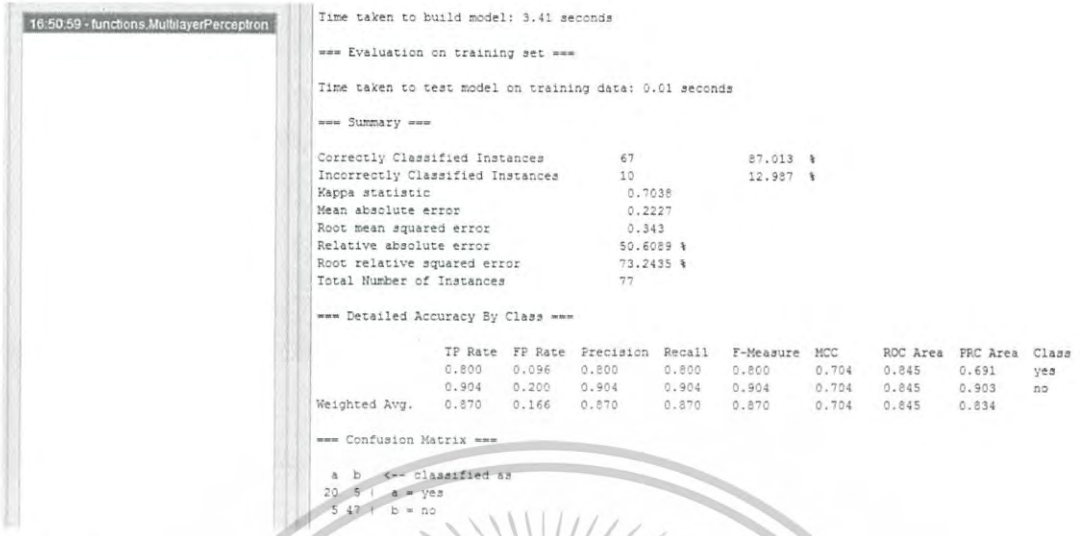
4.5.1.2 โรคเบาหวานของชาวพิม่า ประเทศอินเดีย (Pima Indians Diabetes) จำนวนข้อมูลทดสอบ 768 ค่า ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่ปานกลาง เมื่อกำหนดค่า Random Seed 10 20 และ 30 ตามลำดับ



รูปที่ 4-79 โครงข่ายประสาท โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 10

จากรูป 4-79 โครงข่ายประสาท โรคเบาหวานของชาวพิม่า ประเทศอินเดีย ชั้นข้อมูลเข้า อยู่ด้านซ้ายสุดซึ่งมีคุณลักษณะทั้งหมด 8 ตัวแปร (X_1-X_8) ชั้นซ่อน มี 1 ชั้นซ่อน ชั้นข้อมูลออกหรือชั้นข้อมูลผลลัพธ์ อยู่ด้านขวามือมีคำตอบคือ Yes กับ No Epoch จะรันการทำงานตั้งแต่รอบที่ 1 ถึง 20,000 Error per Epoch แสดงความคลาดเคลื่อนต่อรอบ คือ Learning Rate เท่ากับ 0.1 และ Momentum เท่ากับ 0.9

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4-80 ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 10

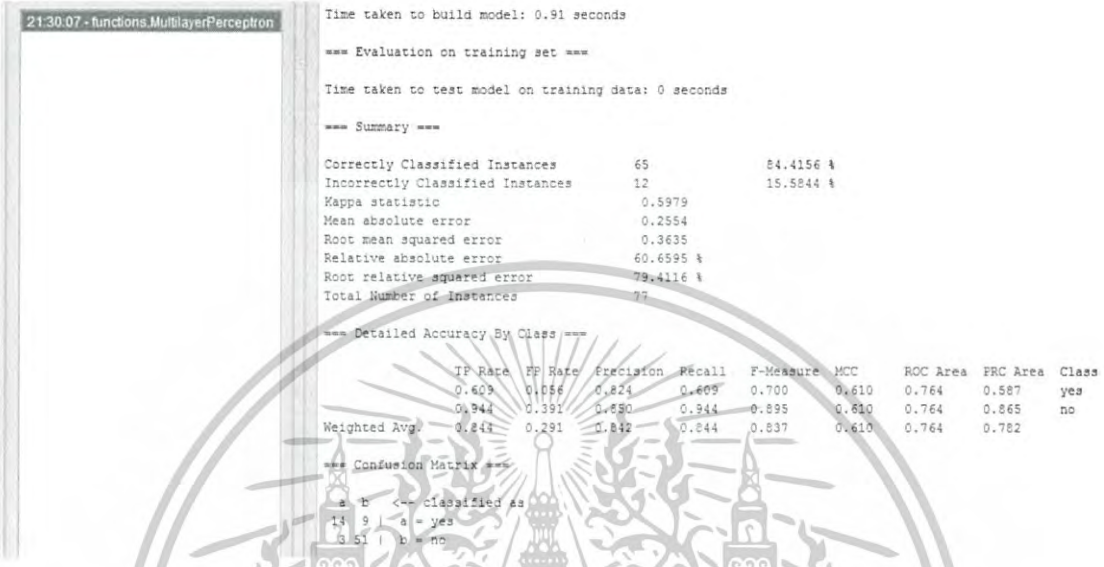
จากรูป 4-80 สุ่มครั้งที่ 1 Random Seed = 10 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 77 คน มีค่าทำนายความถูกต้องจำนวน 67 คน คิดเป็น 87.013% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.343)^2$ = 0.1176 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.2227

รูปที่ 4-81 โครงข่ายประสาท โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 20

จากรูป 4-81 โครงข่ายประสาท โรคเบาหวานของชาวพิม่า ประเทศอินเดีย ชั้นข้อมูลเข้า อยู่ด้านซ้ายสุดซึ่งมีคุณลักษณะทั้งหมด 8 ตัวแปร (X_1-X_8) ชั้นซ่อน มี 1 ชั้นซ่อน ชั้นข้อมูลออกหรือชั้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

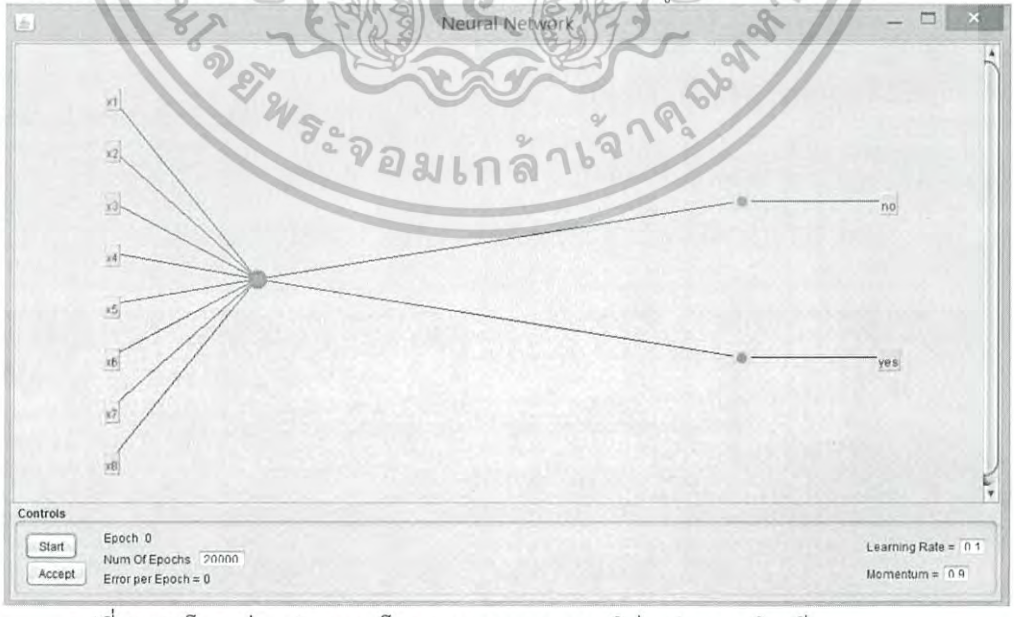
ข้อมูลผลลัพธ์ อยู่ด้านขวามือมีคำตอบคือ Yes กับ No Epoch จะรันการทำงานตั้งแต่รอบที่ 1 ถึง 20,000 Error per Epoch แสดงความคลาดเคลื่อนต่อรอบ คือ Learning Rate เท่ากับ 0.1 และ Momentum เท่ากับ 0.9



```
21:30:07 - functions.MultilayerPerceptron

Time taken to build model: 0.91 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          65               84.4156 %
Incorrectly Classified Instances        12               15.5844 %
Kappa statistic                          0.5979
Mean absolute error                      0.2554
Root mean squared error                  0.3635
Relative absolute error                 60.6595 %
Root relative squared error             79.4116 %
Total Number of Instances               77

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.609    0.056    0.824      0.609   0.700      0.610    0.764     0.587     yes
               0.944    0.391    0.850      0.944   0.895      0.610    0.764     0.865     no
Weighted Avg.  0.844    0.291    0.842      0.844   0.837      0.610    0.764     0.782

=== Confusion Matrix ===

  a  b   <-- classified as
 14  9 |  a = yes
  3 51 |  b = no
```

รูปที่ 4-82 ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 20

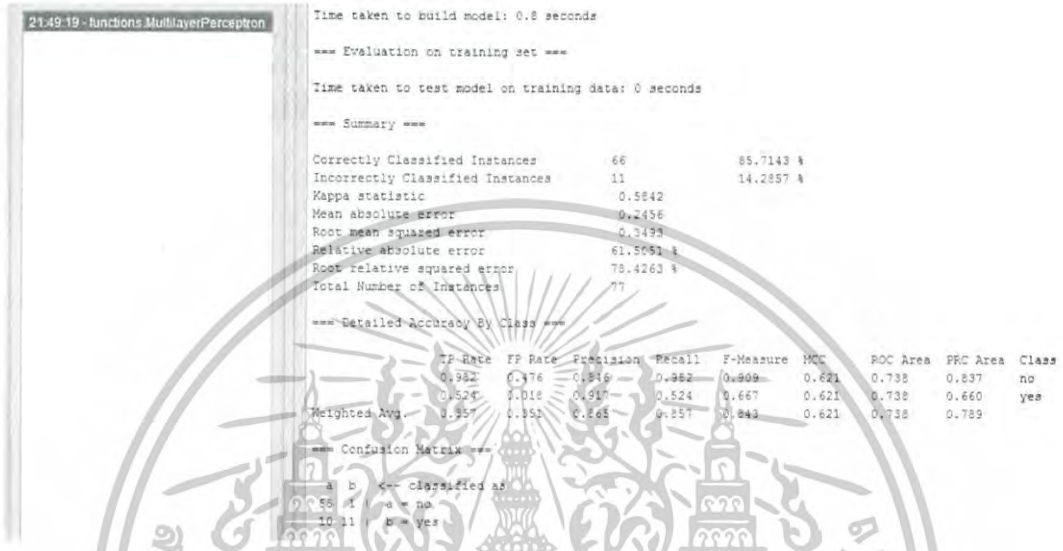
จากรูป 4-82 สุ่มครั้งที่ 2 Random Seed = 20 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 77 คน มีค่าทำนายความถูกต้องจำนวน 65 คน คิดเป็น 84.41560% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.3635)^2$ = 0.1321 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.2554



รูปที่ 4-83 โครงข่ายประสาท โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 30

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากรูป 4-83 โครงข่ายประสาท โรคเบาหวานของชาวพิม่า ประเทศอินเดีย ชั้นข้อมูลเข้า อยู่ด้านซ้ายสุดซึ่งมีคุณลักษณะทั้งหมด 8 ตัวแปร (X_1-X_8) ชั้นซ่อน มี 1 ชั้นซ่อน ชั้นข้อมูลออกหรือชั้นข้อมูลผลลัพธ์ อยู่ด้านขวามือมีคำตอบคือ Yes กับ No Epoch จะรันการทำงานตั้งแต่รอบที่ 1 ถึง 20,000 Error per Epoch แสดงความคลาดเคลื่อนต่อรอบ คือ Learning Rate เท่ากับ 0.1 และ Momentum เท่ากับ 0.9



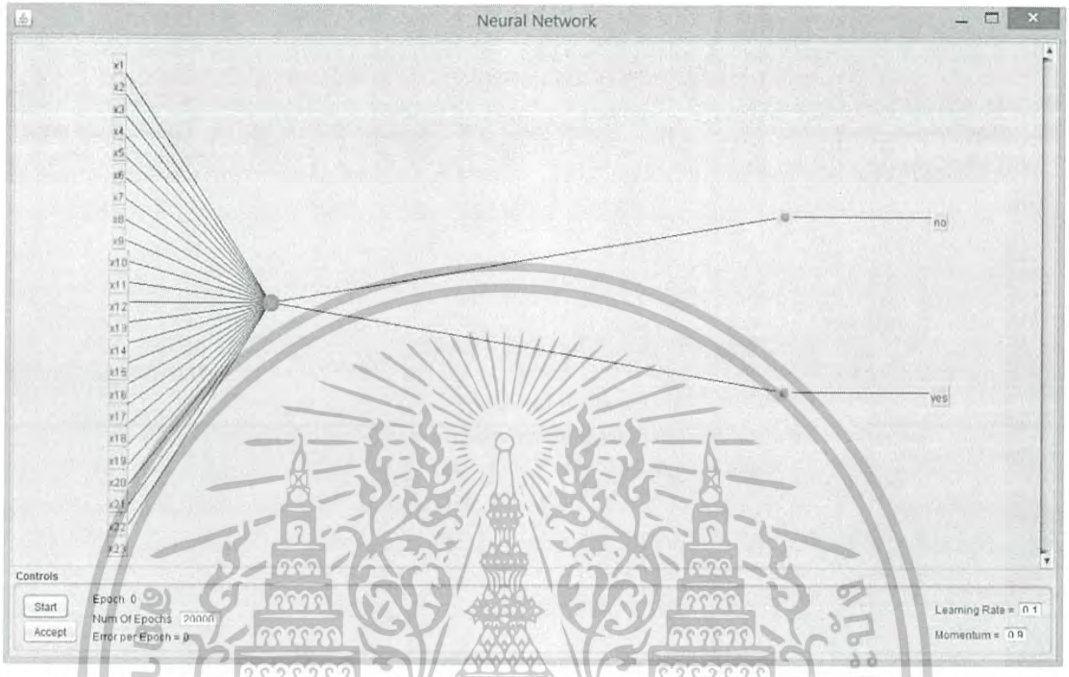
```
21:49:19 - functions.MultilayerPerceptron

Time taken to build model: 0.8 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          66               85.7143 %
Incorrectly Classified Instances        11               14.2857 %
Kappa statistic                          0.5842
Mean absolute error                      0.2456
Root mean squared error                  0.3493
Relative absolute error                 61.5651 %
Root relative squared error             78.4263 %
Total Number of Instances               77

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0.952     0.476     0.846     0.952     0.909     0.621     0.738     0.837     no
                 0.524     0.048     0.917     0.524     0.667     0.621     0.738     0.660     yes
Weighted Avg.    0.857     0.381     0.865     0.857     0.843     0.621     0.738     0.789

=== Confusion Matrix ===

  a  b   <-- classified as
 56  1 |  a = no
 10 11 |  b = yes
```

รูปที่ 4-84 ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 30

จากรูป 4-84 สุ่มครั้งที่ 3 Random Seed = 30 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 77 คน มีค่าทำนายความถูกต้องจำนวน 66 คน คิดเป็น 85.7143% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.3493)^2$ = 0.1220 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.2456

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

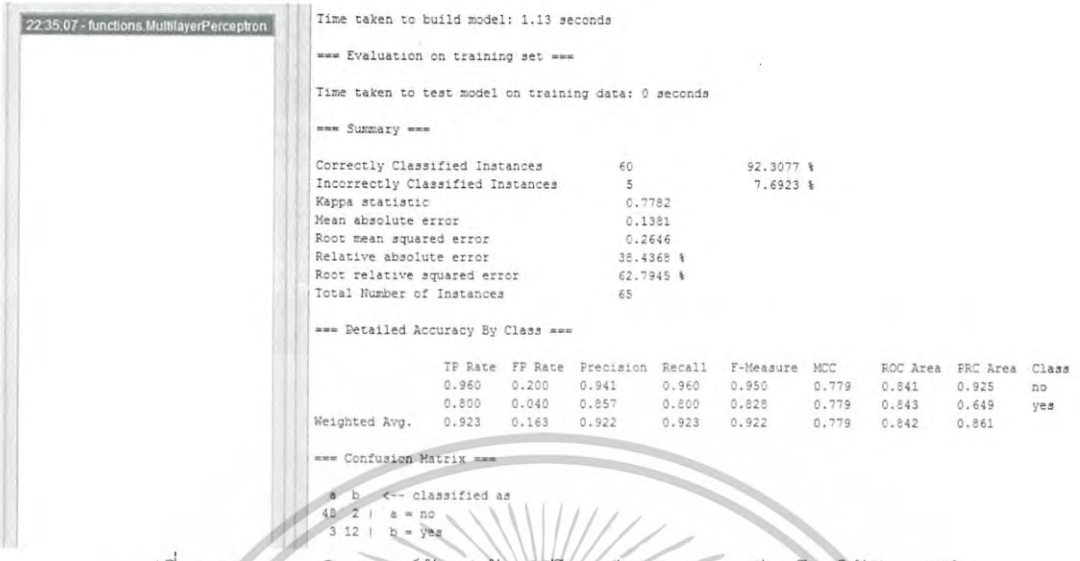
4.5.1.3 การชำระเงินด้วยบัตรเครดิตของลูกค้า (Default of Credit Card Clients) จำนวนข้อมูลทดสอบ 646 ค่า ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่สูง เมื่อกำหนดค่า Random Seed 10 20 และ 30 ตามลำดับ



รูปที่ 4-85 โครงข่ายประสาท การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 10

จากรูป 4-85 โครงข่ายประสาท การชำระเงินด้วยบัตรเครดิตของลูกค้า ชั้นข้อมูลเข้า อยู่ด้านซ้ายสุดซึ่งมีคุณลักษณะทั้งหมด 23 ตัวแปร (X_1-X_{23}) ชั้นซ่อน มี 1 ชั้นซ่อน ชั้นข้อมูลออกหรือชั้นข้อมูลผลลัพธ์ อยู่ด้านขวามือมีคำตอบคือ Yes กับ No Epoch จะรันการทำงานตั้งแต่รอบที่ 1 ถึง 20,000 Error per Epoch แสดงความคลาดเคลื่อนต่อรอบ คือ Learning Rate เท่ากับ 0.1 และ Momentum เท่ากับ 0.9

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

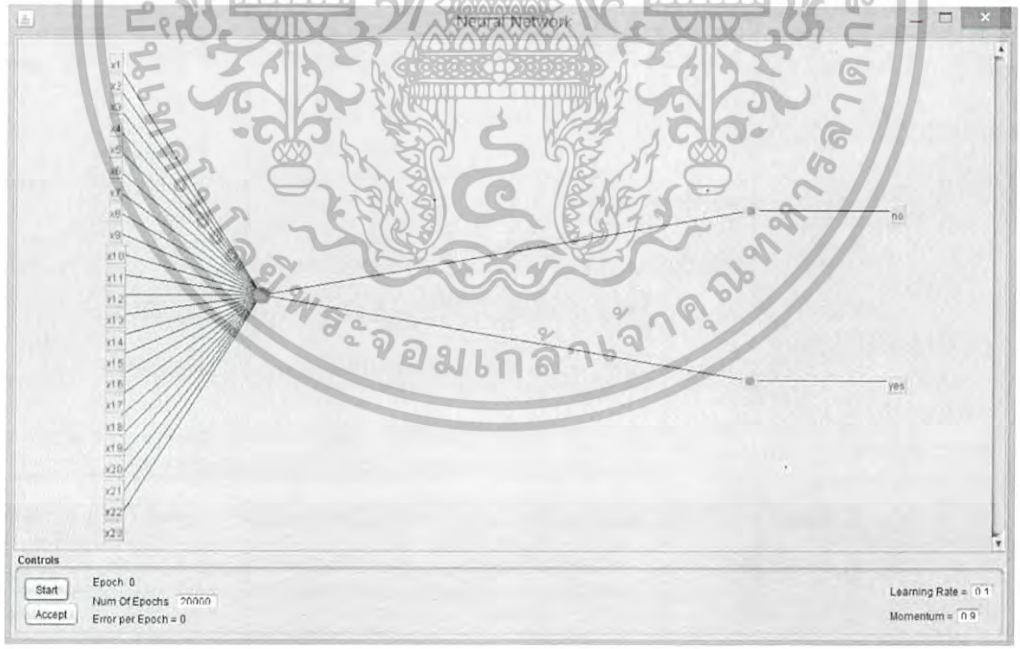


22:35:07 - functions.MultilayerPerceptron

```
Time taken to build model: 1.13 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          60               92.3077 %
Incorrectly Classified Instances         5                7.6923 %
Kappa statistic                          0.7782
Mean absolute error                      0.1381
Root mean squared error                  0.2646
Relative absolute error                 38.4368 %
Root relative squared error             62.7945 %
Total Number of Instances               65

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.960    0.200    0.941      0.960   0.950      0.779    0.841     0.925     no
               0.800    0.040    0.857      0.800   0.828      0.779    0.843     0.649     yes
Weighted Avg.  0.923    0.163    0.922      0.923   0.922      0.779    0.842     0.861

=== Confusion Matrix ===

  a  b   <-- classified as
 48  2 |  a = no
  3 12 |  b = yes
```

รูปที่ 4-86 ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 10

จากรูป 4-86 สุ่มครั้งที่ 1 Random Seed = 10 กลุ่มลูกค้า 10 เปอร์เซ็นต์ จำนวนทั้งหมด 65 คน มีค่าทำนายความถูกต้องจำนวน 60 คน คิดเป็น 92.3077% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.2646)^2$ = 0.0700 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.1381

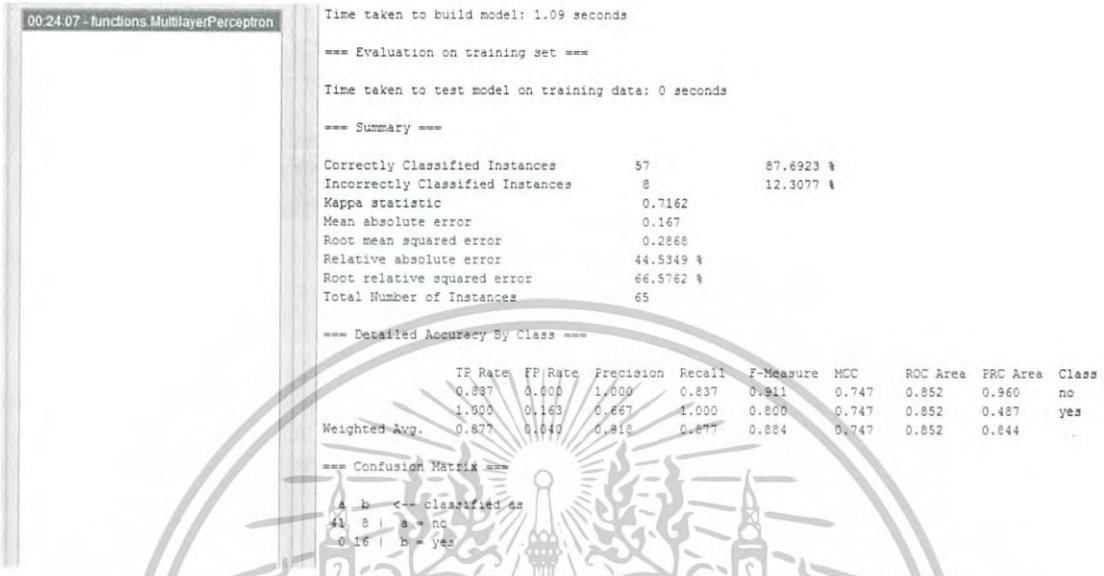


รูปที่ 4-87 โครงข่ายประสาท การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 20

จากรูป 4-87 โครงข่ายประสาท การชำระเงินด้วยบัตรเครดิตของลูกค้า ชั้นข้อมูลเข้า อยู่ด้านซ้ายสุดซึ่งมีคุณลักษณะทั้งหมด 23 ตัวแปร (X_1-X_{23}) ชั้นซ่อน มี 1 ชั้นซ่อน ชั้นข้อมูลออกหรือชั้นข้อมูลผลลัพธ์ อยู่ด้านขวามือมีคำตอบคือ Yes กับ No Epoch จะรันการทำงานตั้งแต่รอบที่ 1 ถึง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

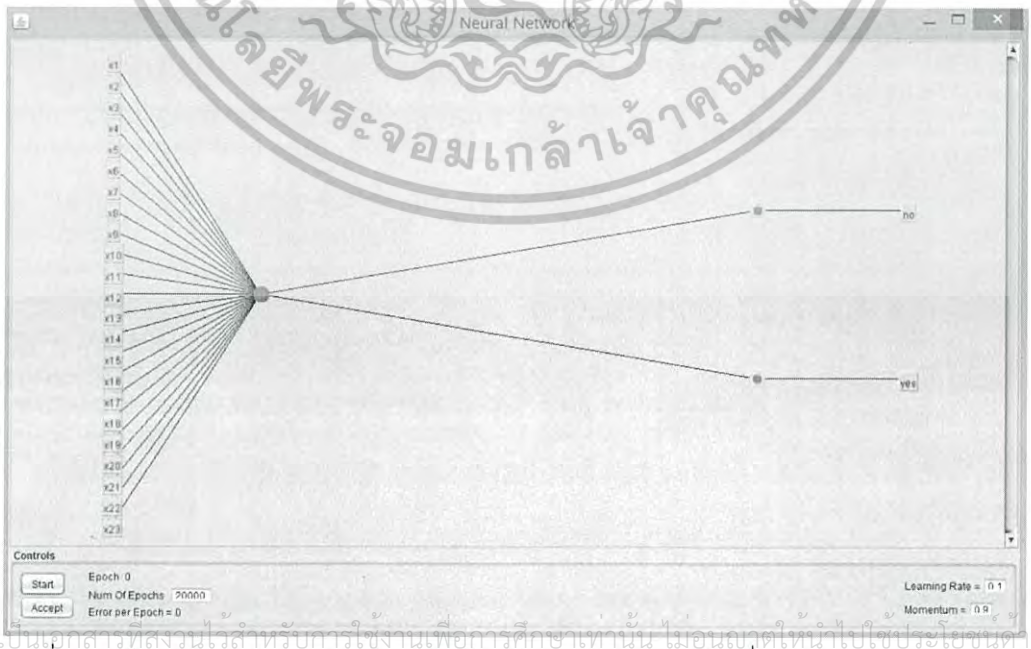
20,000 Error per Epoch แสดงความคลาดเคลื่อนต่อรอบ คือ Learning Rate เท่ากับ 0.1 และ Momentum เท่ากับ 0.9



```
00:24:07 - functions.MultilayerPerceptron

Time taken to build model: 1.09 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          57               87.6923 %
Incorrectly Classified Instances         8               12.3077 %
Kappa statistic                          0.7162
Mean absolute error                      0.167
Root mean squared error                  0.2868
Relative absolute error                 44.5349 %
Root relative squared error             66.5762 %
Total Number of Instances               65

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.837    0.000    1.000      0.837   0.911      0.747    0.852     0.960     no
                 1.000    0.163    0.667      1.000   0.800      0.747    0.852     0.487     yes
Weighted Avg.    0.877    0.040    0.918      0.877   0.884      0.747    0.852     0.844

=== Confusion Matrix ===

  a  b   <-- classified as
 41  8 |  a = no
  0 16 |  b = yes
```

รูปที่ 4-88 ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 20

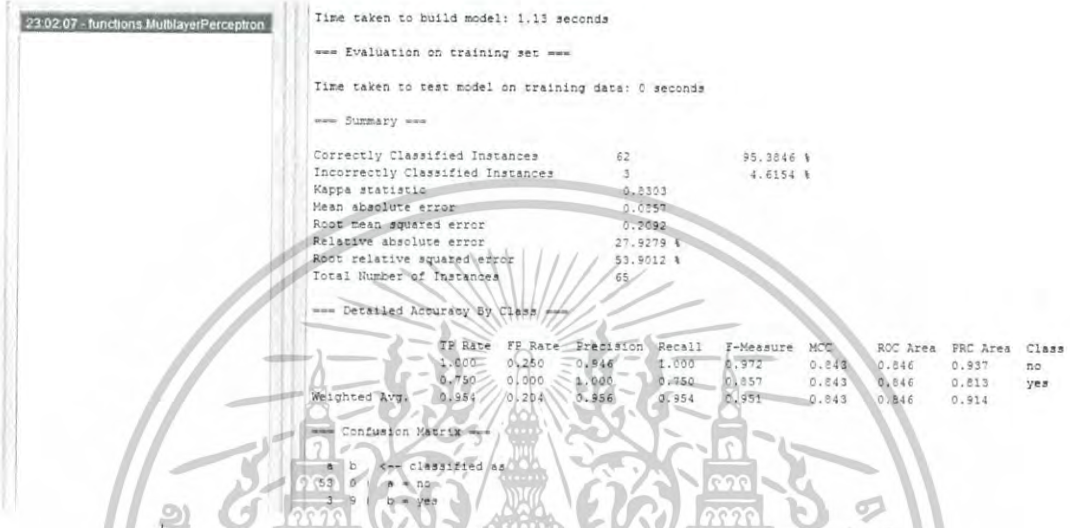
จากรูป 4-88 สุ่มครั้งที่ 2 Random Seed = 20 กลุ่มลูกค้า 10 เปอร์เซ็นต์ จำนวนทั้งหมด 65 คน มีค่าทำนายความถูกต้องจำนวน 57 คน คิดเป็น 87.6923% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.2868)^2$ = 0.0823 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.167



รูปที่ 4-89 โครงข่ายประสาท การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 30

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากรูป 4-89 โครงข่ายประสาท การชำระเงินด้วยบัตรเครดิตของลูกค้า ชั้นข้อมูลเข้า อยู่ด้านซ้ายสุดซึ่งมีคุณลักษณะทั้งหมด 23 ตัวแปร (X_1-X_{23}) ชั้นซ่อน มี 1 ชั้นซ่อน ชั้นข้อมูลออกหรือชั้นข้อมูลผลลัพธ์ อยู่ด้านขวามือมีคำตอบคือ Yes กับ No Epoch จะรันการทำงานตั้งแต่รอบที่ 1 ถึง 20,000 Error per Epoch แสดงความคลาดเคลื่อนต่อรอบ คือ Learning Rate เท่ากับ 0.1 และ Momentum เท่ากับ 0.9



23:02:07 - functions.MultilayerPerceptron

```
Time taken to build model: 1.13 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          62               95.3846 %
Incorrectly Classified Instances         3                4.6154 %
Kappa statistic                          0.8303
Mean absolute error                      0.0857
Root mean squared error                  0.2092
Relative absolute error                 27.9279 %
Root relative squared error             53.9012 %
Total Number of Instances               65

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.250      0.946     1.000     0.972     0.843     0.846     0.937    no
                 0.750    0.000      1.000     0.750     0.857     0.843     0.846     0.813    yes
Weighted Avg.    0.954    0.204      0.956     0.954     0.951     0.843     0.846     0.914

=== Confusion Matrix ===

  a  b   <-- classified as
 53  0 |  a = no
  3  9 |  b = yes
```

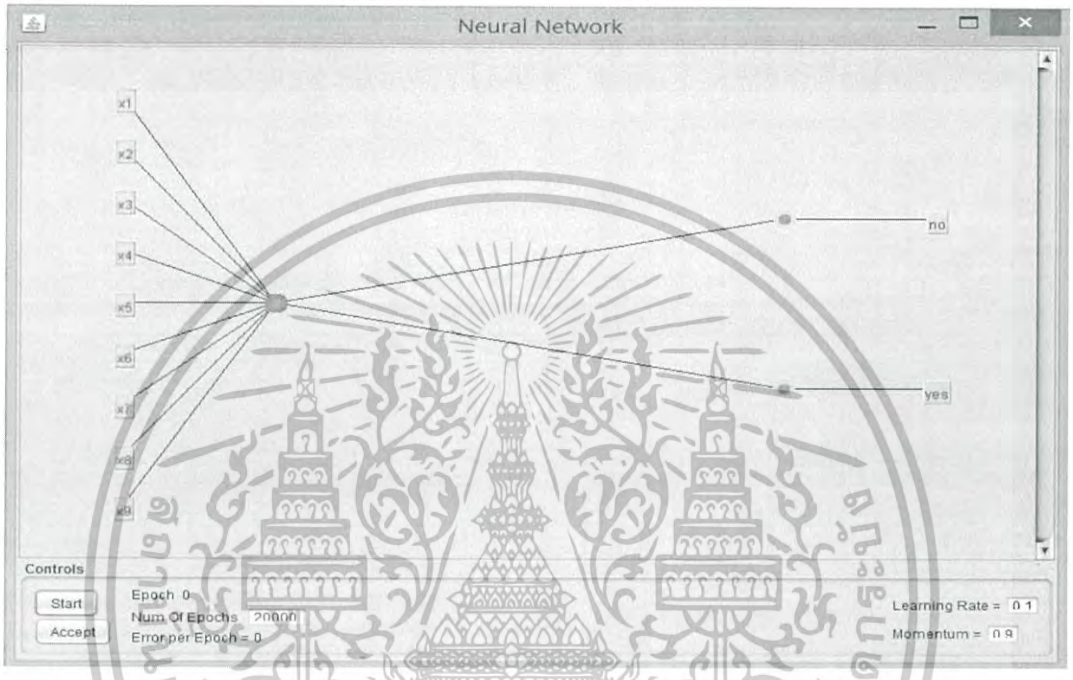
รูปที่ 4-90 ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 30

จากรูป 4-90 สุ่มครั้งที่ 3 Random Seed = 30 กลุ่มลูกค้า 10 เปอร์เซ็นต์ จำนวนทั้งหมด 65 คน มีค่าทำนายความถูกต้องจำนวน 62 คน คิดเป็น 95.3846% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.2092)^2$ = 0.0438 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0857

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.5.2 ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron โดยโปรแกรม WEKA

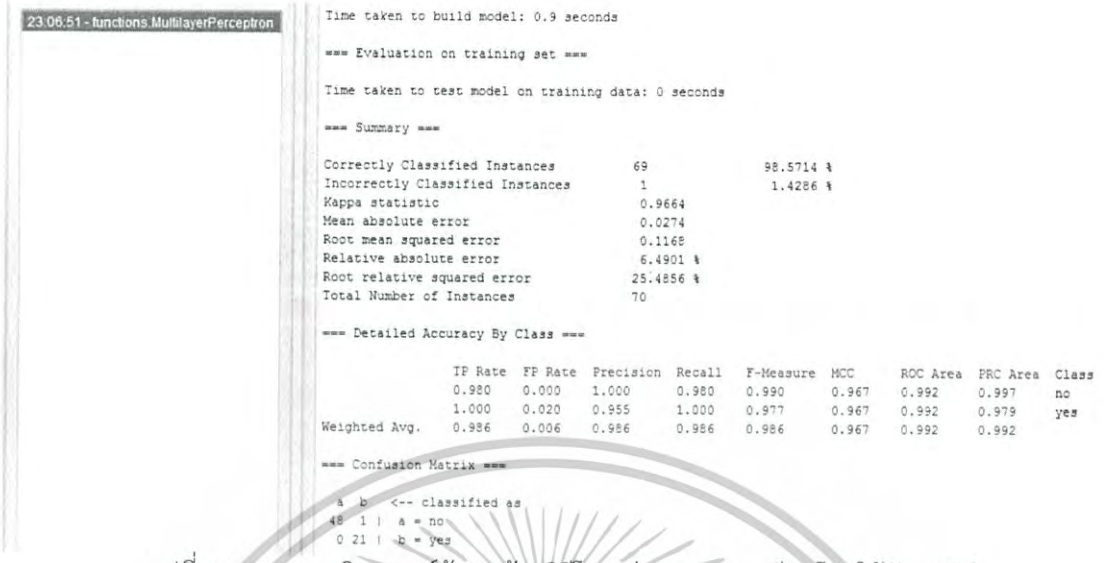
4.5.2.1 โรคมะเร็งเต้านมของรัฐวิสคอนซิน (Breast Cancer Wisconsin) จำนวนข้อมูลทดสอบ 699 ค่า ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่ต่ำ เมื่อกำหนดค่า Random Seed 10 20 และ 30 ตามลำดับ



รูปที่ 4-91 โครงข่ายประสาท โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 10

จากรูป 4-91 โครงข่ายประสาท โรคมะเร็งเต้านมของรัฐวิสคอนซิน ชั้นข้อมูลเข้า อยู่ด้านซ้ายสุดซึ่งมีคุณลักษณะทั้งหมด 9 ตัวแปร (X_1-X_9) ชั้นซ่อน มี 1 ชั้นซ่อน ชั้นข้อมูลออกหรือชั้นข้อมูลผลลัพธ์ อยู่ด้านขวามือมีคำตอบคือ Yes กับ No Epoch จะรันการทำงานตั้งแต่รอบที่ 1 ถึง 20,000 Error per Epoch แสดงความคลาดเคลื่อนต่อรอบ คือ Learning Rate เท่ากับ 0.1 และ Momentum เท่ากับ 0.9

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



```
23:06:51 - functions.MultilayerPerceptron

Time taken to build model: 0.9 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          69               98.5714 %
Incorrectly Classified Instances         1                1.4286 %
Kappa statistic                          0.9664
Mean absolute error                      0.0274
Root mean squared error                  0.1168
Relative absolute error                  6.4901 %
Root relative squared error             25.4856 %
Total Number of Instances               70

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.980    0.000    1.000      0.980    0.990      0.967    0.992     0.997     no
                 1.000    0.020    0.955      1.000    0.977      0.967    0.992     0.979     yes
Weighted Avg.    0.986    0.006    0.986      0.986    0.986      0.967    0.992     0.992

=== Confusion Matrix ===

  a  b   <-- classified as
 48  1 |  a = no
  0 21 |  b = yes
```

รูปที่ 4-92 ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 10

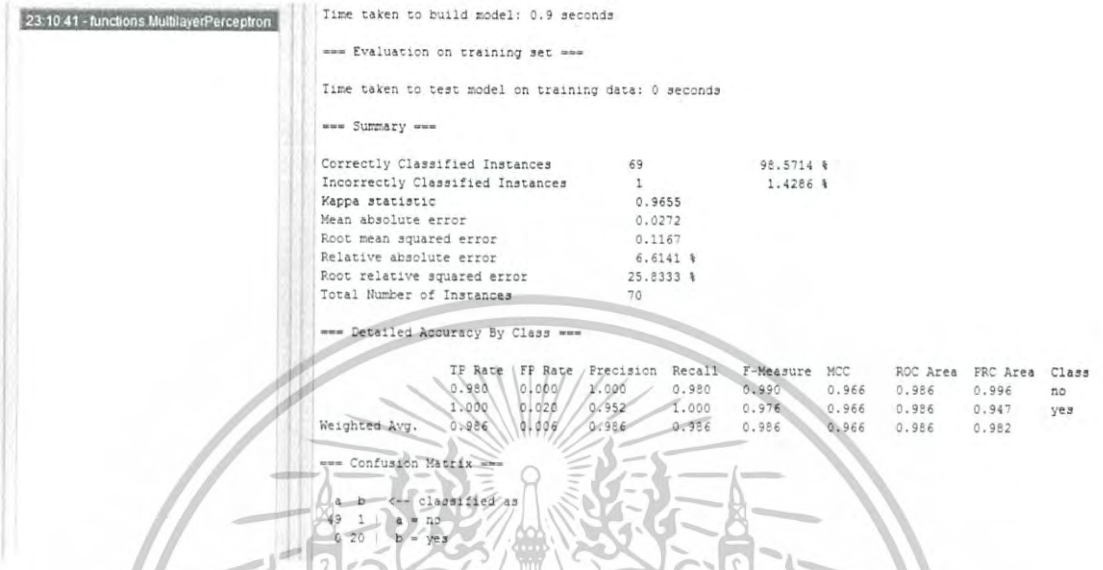
จากรูป 4-92 สุ่มครั้งที่ 1 Random Seed = 10 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 70 คน มีค่าทำนายความถูกต้องจำนวน 69 คน คิดเป็น 98.5714% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.1168)^2$ = 0.0136 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0274

รูปที่ 4-93 โครงข่ายประสาท โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 20

จากรูป 4-93 โครงข่ายประสาท โรคมะเร็งเต้านมของรัฐวิสคอนซิน ชั้นข้อมูลเข้า อยู่ด้านซ้ายสุดซึ่งมีคุณลักษณะทั้งหมด 9 ตัวแปร (X_1-X_9) ชั้นซ่อน มี 1 ชั้นซ่อน ชั้นข้อมูลออกหรือชั้นข้อมูลผลลัพธ์ อยู่ด้านขวามือมีคำตอบคือ Yes กับ No Epoch จะรันการทำงานตั้งแต่รอบที่ 1 ถึง 20,000

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Error per Epoch แสดงความคลาดเคลื่อนต่อรอบ คือ Learning Rate เท่ากับ 0.1 และ Momentum เท่ากับ 0.9

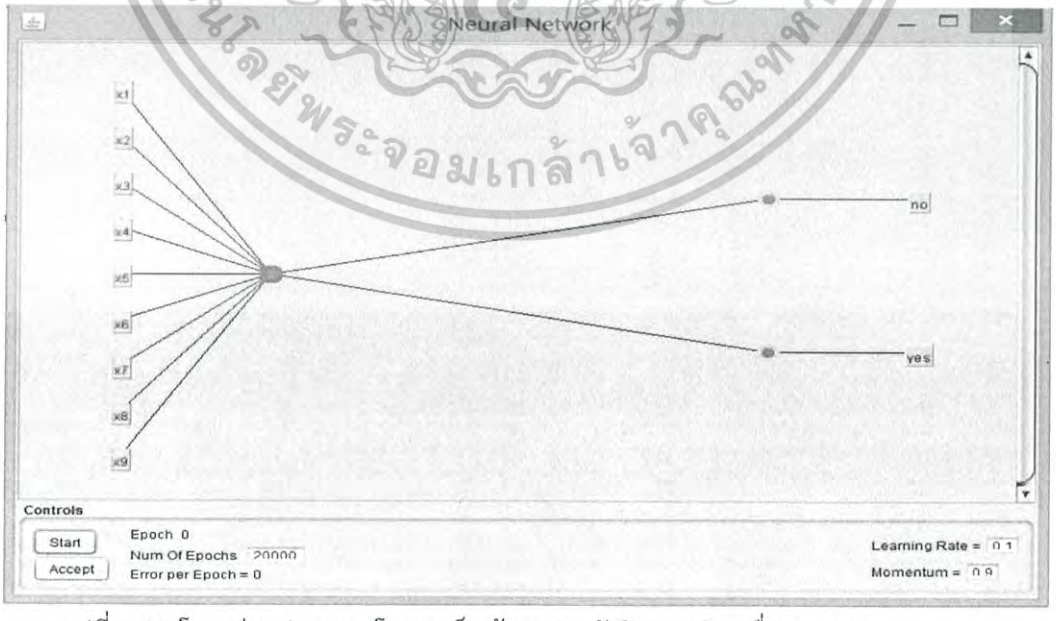


23:10:41 - functions.MultilayerPerceptron

```
Time taken to build model: 0.9 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          69               98.5714 %
Incorrectly Classified Instances         1                1.4286 %
Kappa statistic                          0.9655
Mean absolute error                      0.0272
Root mean squared error                  0.1167
Relative absolute error                  6.6141 %
Root relative squared error             25.8333 %
Total Number of Instances               70

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0.980     0.000     1.000     0.980     0.990     0.966     0.986     0.996     no
                 1.000     0.020     0.952     1.000     0.976     0.966     0.986     0.947     yes
Weighted Avg.    0.986     0.006     0.986     0.986     0.986     0.966     0.986     0.982

=== Confusion Matrix ===

  a  b   <-- classified as
 49  1 |  a = no
  0 20 |  b = yes
```

รูปที่ 4-94 ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 20

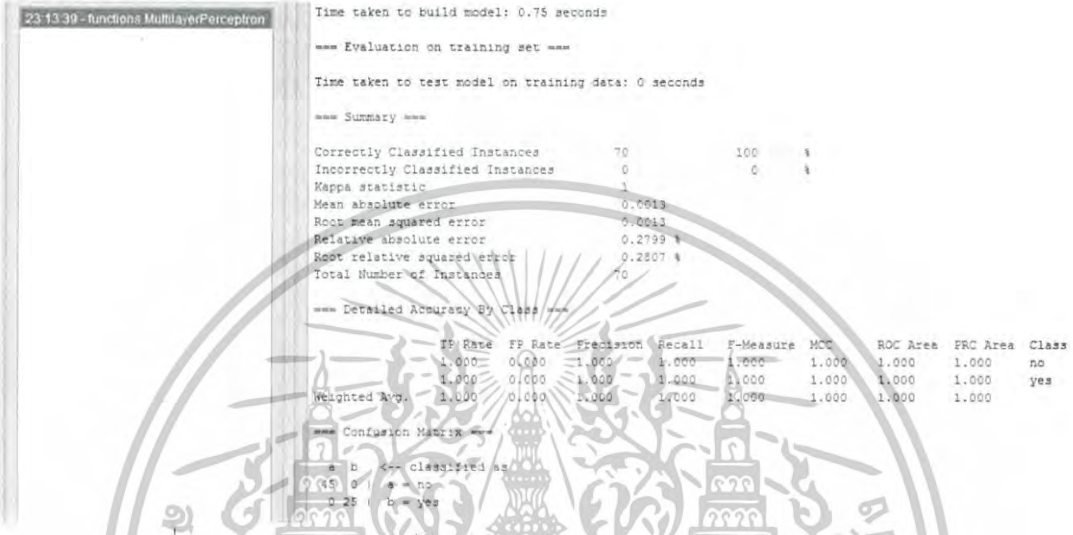
จากรูป 4-94 สุ่มครั้งที่ 2 Random Seed = 20 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 70 คน มีค่าทำนายความถูกต้องจำนวน 69 คน คิดเป็น 98.5714% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.1167)^2$ = 0.01362 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0272



รูปที่ 4-95 โครงข่ายประสาท โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 30

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากรูป 4-95 โครงข่ายประสาท โรคมะเร็งเต้านมของรัฐวิสคอนซิน ชั้นข้อมูลเข้า อยู่ด้านซ้ายสุดซึ่งมีคุณลักษณะทั้งหมด 9 ตัวแปร (X_1-X_9) ชั้นซ่อน มี 1 ชั้นซ่อน ชั้นข้อมูลออกหรือชั้นข้อมูลผลลัพธ์ อยู่ด้านขวามือมีคำตอบคือ Yes กับ No Epoch จะรันการทำงานตั้งแต่รอบที่ 1 ถึง 20,000 Error per Epoch แสดงความคลาดเคลื่อนต่อรอบ คือ Learning Rate เท่ากับ 0.1 และ Momentum เท่ากับ 0.9

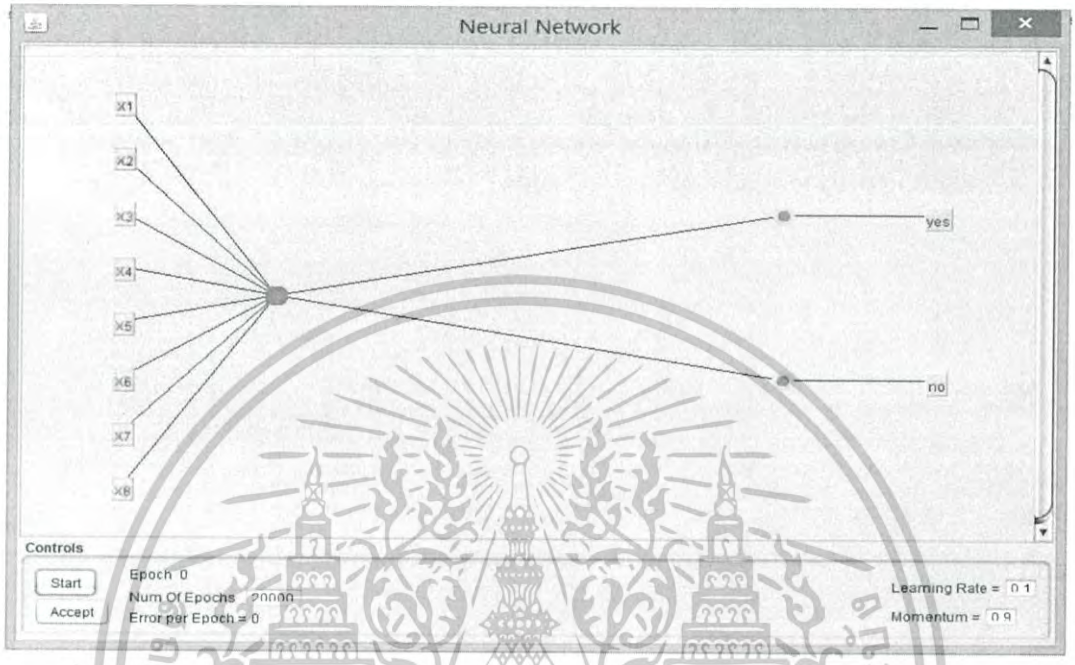


รูปที่ 4-96 ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 30

จากรูป 4-96 สุ่มครั้งที่ 3 Random Seed = 30 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 70 คน มีค่าทำนายความถูกต้องจำนวน 70 คน คิดเป็น 100% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.0013)^2$ = 0.0000 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0013

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

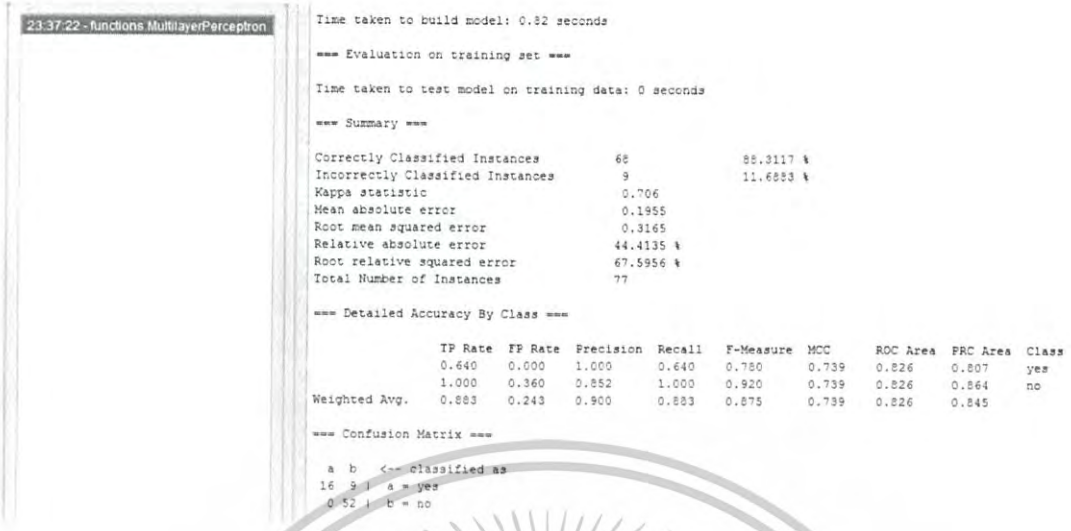
4.5.2.2 โรคเบาหวานของชาวพิม่า ประเทศอินเดีย (Pima Indians Diabetes) จำนวนข้อมูลทดสอบ 768 ค่า ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่ปานกลาง เมื่อกำหนดค่า Random Seed 10 20 และ 30 ตามลำดับ



รูปที่ 4-97 โครงข่ายประสาท โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 10

จากรูป 4-97 โครงข่ายประสาท โรคมะเร็งเต้านมของรัฐวิสคอนซิน ชั้นข้อมูลเข้า อยู่ด้านซ้ายสุดซึ่งมีคุณลักษณะทั้งหมด 8 ตัวแปร (X_1-X_8) ชั้นซ่อน มี 1 ชั้นซ่อน ชั้นข้อมูลออกหรือชั้นข้อมูลผลลัพธ์ อยู่ด้านขวามือมีคำตอบคือ Yes กับ No Epoch จะรันการทำงานตั้งแต่รอบที่ 1 ถึง 20,000 Error per Epoch แสดงความคลาดเคลื่อนต่อรอบ คือ Learning Rate เท่ากับ 0.1 และ Momentum เท่ากับ 0.9

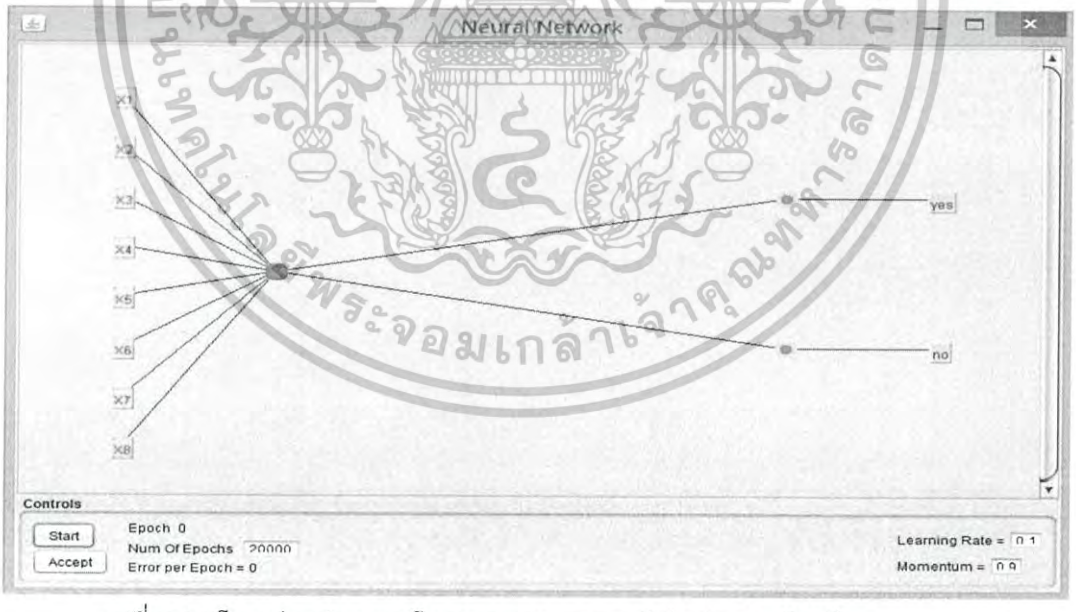
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



```
23:37:22 - functions.MultilayerPerceptron

Time taken to build model: 0.82 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          68               88.3117 %
Incorrectly Classified Instances         9               11.6883 %
Kappa statistic                          0.706
Mean absolute error                      0.1955
Root mean squared error                  0.3165
Relative absolute error                 44.4135 %
Root relative squared error             67.5956 %
Total Number of Instances               77

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.640     0.000     1.000       0.640   0.780      0.739    0.826     0.807     yes
               1.000     0.360     0.852       1.000   0.920      0.739    0.826     0.864     no
Weighted Avg.  0.883     0.243     0.900       0.883   0.875      0.739    0.826     0.845

=== Confusion Matrix ===

  a  b   <-- classified as
 16  9 |  a = yes
  0 52 |  b = no
```

รูปที่ 4-98 ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 10

จากรูป 4-98 สุ่มครั้งที่ 1 Random Seed = 10 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 77 คน มีค่าทำนายความถูกต้องจำนวน 68 คน คิดเป็น 88.3117% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.3165)^2$ = 0.1002 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.1955

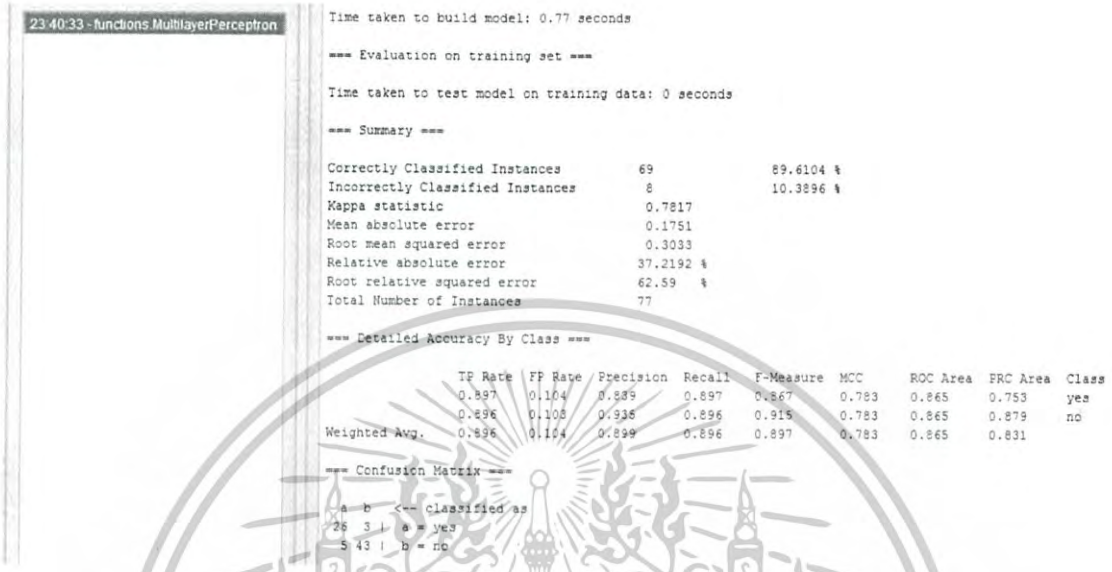


รูปที่ 4-99 โครงข่ายประสาท โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 20

จากรูป 4-99 โครงข่ายประสาท โรคมะเร็งเต้านมของรัฐวิสคอนซิน ชั้นข้อมูลเข้า อยู่ด้านซ้ายสุดซึ่งมีคุณลักษณะทั้งหมด 8 ตัวแปร (X_1-X_8) ชั้นซ่อน มี 1 ชั้นซ่อน ชั้นข้อมูลออกหรือชั้นข้อมูลผลลัพธ์ อยู่ด้านขวามือมีคำตอบคือ Yes กับ No Epoch จะรันการทำงานตั้งแต่รอบที่ 1 ถึง 20,000

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Error per Epoch แสดงความคลาดเคลื่อนต่อรอบ คือ Learning Rate เท่ากับ 0.1 และ Momentum เท่ากับ 0.9

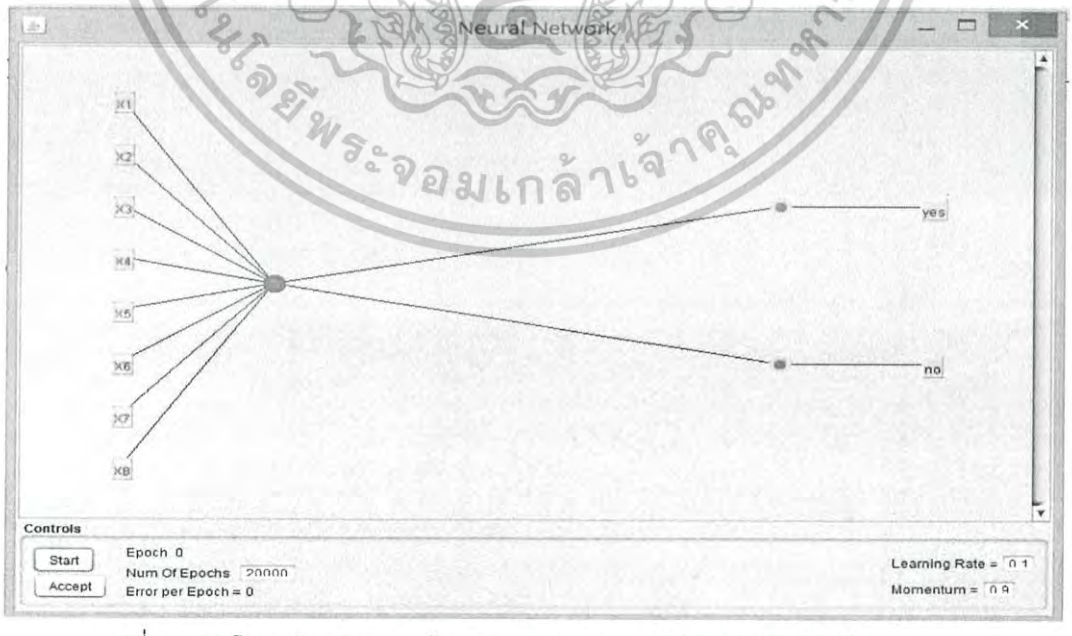


23:40:33 - functions.MultilayerPerceptron

```
Time taken to build model: 0.77 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          69               89.6104 %
Incorrectly Classified Instances         8               10.3896 %
Kappa statistic                          0.7817
Mean absolute error                      0.1751
Root mean squared error                  0.3033
Relative absolute error                 37.2192 %
Root relative squared error             62.59   %
Total Number of Instances               77

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0.897     0.104      0.839     0.897      0.867     0.783     0.865     0.753     yes
                 0.896     0.103      0.935     0.896      0.915     0.783     0.865     0.879     no
Weighted Avg.    0.896     0.104      0.899     0.896      0.897     0.783     0.865     0.831

=== Confusion Matrix ===

  a  b   <-- classified as
 26  3 |  a = yes
  5 43 |  b = no
```

รูปที่ 4-100 ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 20

จากรูป 4-100 สุ่มครั้งที่ 2 Random Seed = 20 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 77 คน มีค่าทำนายความถูกต้องจำนวน 69 คน คิดเป็น 89.6104% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.3033)^2$ = 0.0910 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.1751



รูปที่ 4-101 โครงข่ายประสาท โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 30

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

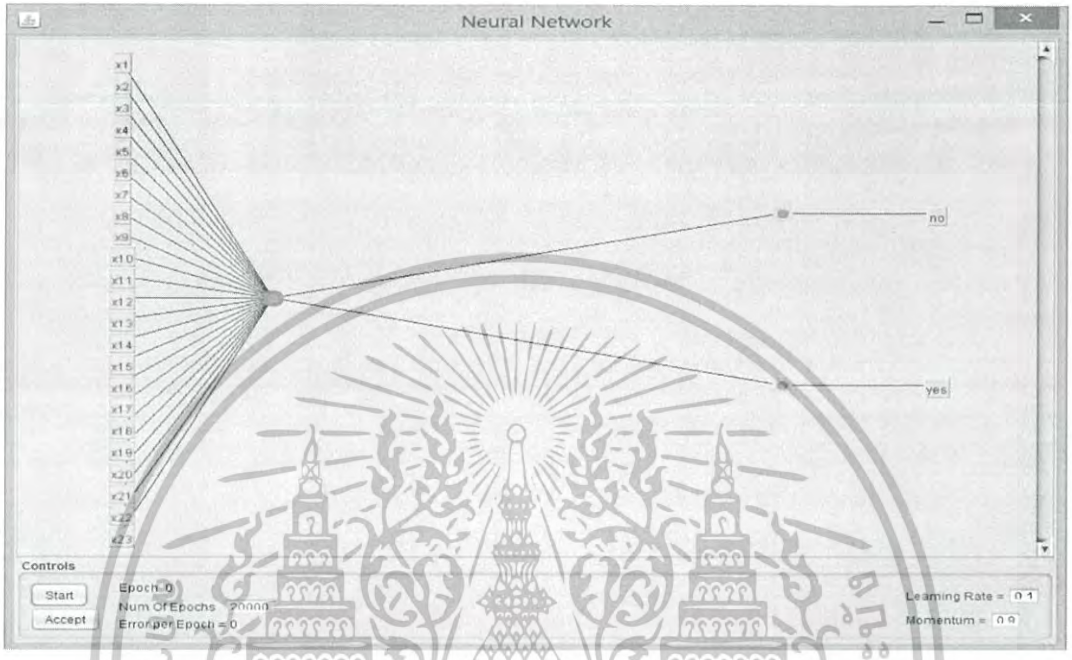
จากรูป 4-101 โครงข่ายประสาท โรคมะเร็งเต้านมของรัฐวิสคอนซิน ชั้นข้อมูลเข้า อยู่ด้านซ้ายสุดซึ่งมีคุณลักษณะทั้งหมด 8 ตัวแปร (X_1-X_8) ชั้นซ่อน มี 1 ชั้นซ่อน ชั้นข้อมูลออกหรือชั้นข้อมูลผลลัพธ์ อยู่ด้านขวามือมีคำตอบคือ Yes กับ No Epoch จะรันการทำงานตั้งแต่รอบที่ 1 ถึง 20,000 Error per Epoch แสดงความคลาดเคลื่อนต่อรอบ คือ Learning Rate เท่ากับ 0.1 และ Momentum เท่ากับ 0.9

รูปที่ 4-102 ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 30

จากรูป 4-102 สุ่มครั้งที่ 3 Random Seed = 30 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 77 คน มีค่าทำนายความถูกต้องจำนวน 71 คน คิดเป็น 92.2078% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.2648)^2$ = 0.0701 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.1387

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.5.2.3 การชำระเงินด้วยบัตรเครดิตของลูกค้า (Default of Credit Card Clients) จำนวนข้อมูลทดสอบ 646 ค่า ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่สูง เมื่อกำหนดค่า Random Seed 10 20 และ 30 ตามลำดับ



รูปที่ 4-103 โครงข่ายประสาท การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 10

จากรูป 4-103 โครงข่ายประสาท โรคมะเร็งเต้านมของรัฐวิสคอนซิน ชั้นข้อมูลเข้า อยู่ด้านซ้ายสุดซึ่งมีคุณลักษณะทั้งหมด 23 ตัวแปร (X_1-X_{23}) ชั้นซ่อน มี 1 ชั้นซ่อน ชั้นข้อมูลออกหรือชั้นข้อมูลผลลัพธ์ อยู่ด้านขวามือมีคำตอบคือ Yes กับ No Epoch จะรันการทำงานตั้งแต่รอบที่ 1 ถึง 20,000 Error per Epoch แสดงความคลาดเคลื่อนต่อรอบ คือ Learning Rate เท่ากับ 0.1 และ Momentum เท่ากับ 0.9

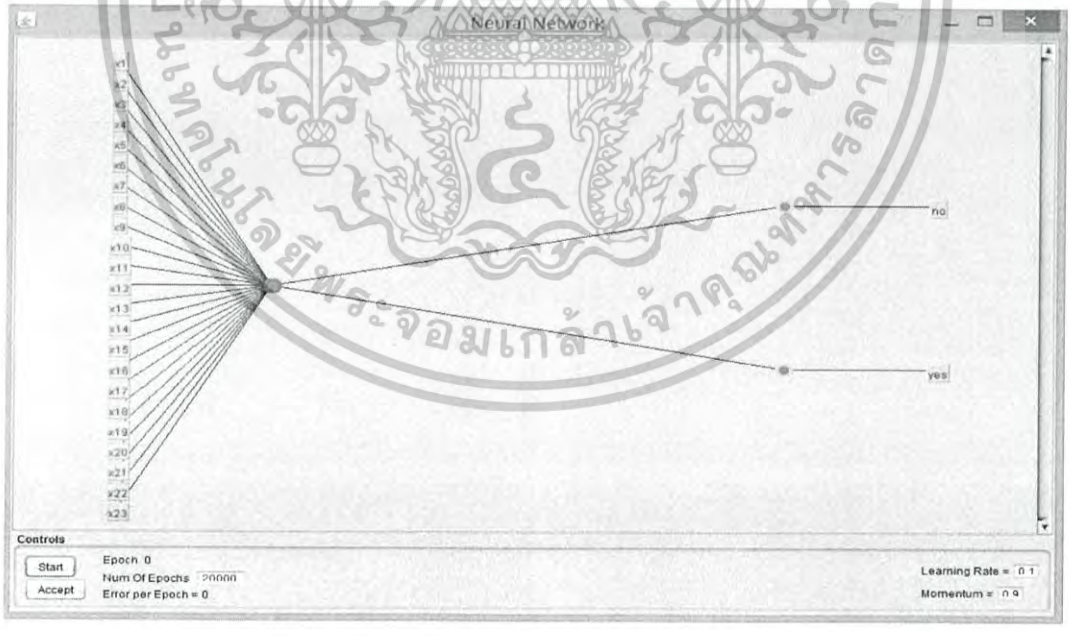
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

23:52:20 - functions.MultilayerPerceptron

```
Time taken to build model: 1.07 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          61               93.8462 %
Incorrectly Classified Instances         4                6.1538 %
Kappa statistic                          0.8471
Mean absolute error                      0.1113
Root mean squared error                  0.2378
Relative absolute error                 25.9841 %
Root relative squared error             51.5246 %
Total Number of Instances               65

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.200    0.918      1.000   0.957      0.857    0.871     0.914     no
                 0.800    0.000    1.000      0.800   0.889      0.857    0.871     0.879     yes
Weighted Avg.    0.938    0.138    0.943      0.938   0.936      0.857    0.871     0.903

=== Confusion Matrix ===

  a  b   <-- classified as
 45  0 |  a = no
  4 16 |  b = yes
```

รูปที่ 4-104 ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 10

จากรูป 4-104 สุ่มครั้งที่ 1 Random Seed = 10 กลุ่มลูกค้า 10 เปอร์เซ็นต์ จำนวนทั้งหมด 65 คน มีค่าทำนายความถูกต้องจำนวน 61 คน คิดเป็น 93.8462% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.2378)^2$ = 0.0566 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.1113

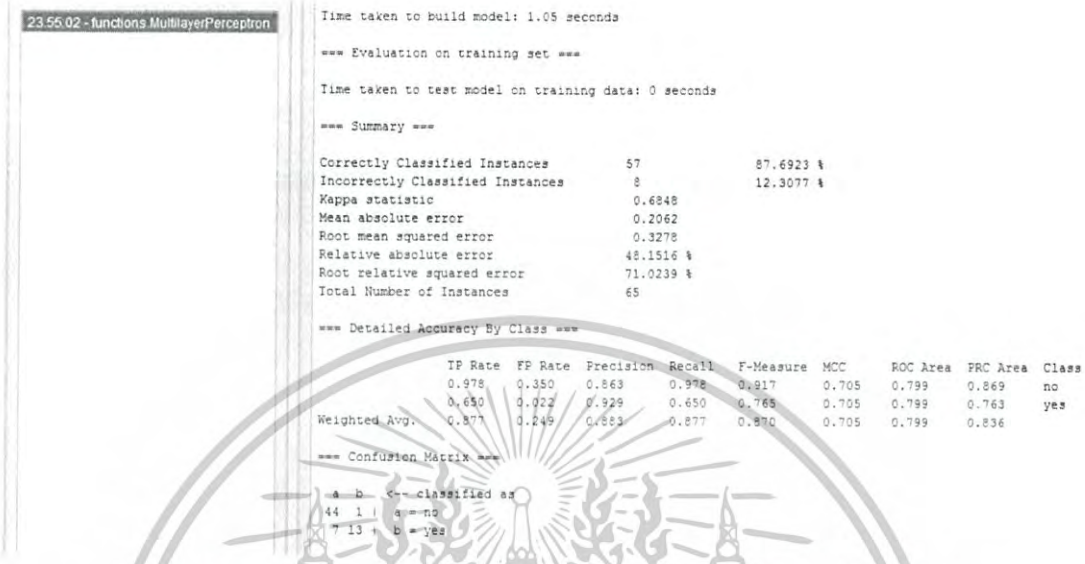


รูปที่ 4-105 โครงข่ายประสาท การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 20

จากรูป 4-105 โครงข่ายประสาท โรคมะเร็งเต้านมของรัฐวิสคอนซิน ชั้นข้อมูลเข้า อยู่ด้านซ้ายสุดซึ่งมีคุณลักษณะทั้งหมด 23 ตัวแปร (X_1-X_{23}) ชั้นซ่อน มี 1 ชั้นซ่อน ชั้นข้อมูลออกหรือชั้นข้อมูลผลลัพธ์ อยู่ด้านขวามือมีคำตอบคือ Yes กับ No Epoch จะรันการทำงานตั้งแต่รอบที่ 1 ถึง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

20,000 Error per Epoch แสดงความคลาดเคลื่อนต่อรอบ คือ Learning Rate เท่ากับ 0.1 และ Momentum เท่ากับ 0.9



23:55:02 - functions.MultilayerPerceptron

Time taken to build model: 1.05 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances	57	87.6923 %
Incorrectly Classified Instances	8	12.3077 %
Kappa statistic	0.6848	
Mean absolute error	0.2062	
Root mean squared error	0.3278	
Relative absolute error	48.1516 %	
Root relative squared error	71.0239 %	
Total Number of Instances	65	

=== Detailed Accuracy By Class ===

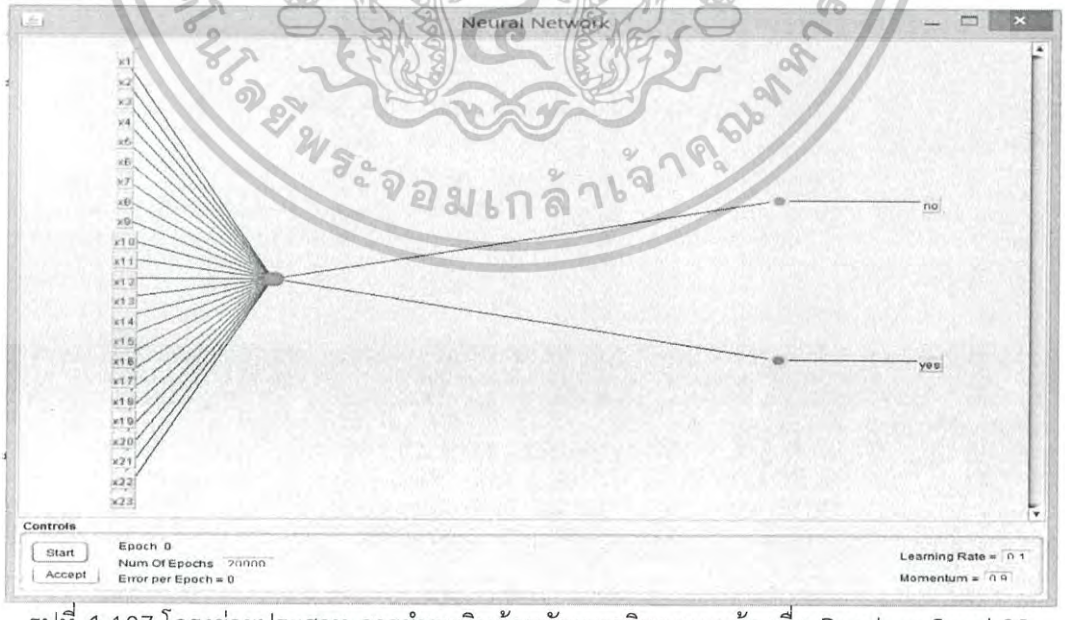
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.978	0.350	0.863	0.978	0.917	0.705	0.799	0.869	no
	0.650	0.022	0.929	0.650	0.765	0.705	0.799	0.763	yes
Weighted Avg.	0.877	0.249	0.883	0.877	0.870	0.705	0.799	0.836	

=== Confusion Matrix ===

a	b	<-- classified as
44	1	a = no
7	13	b = yes

รูปที่ 4-106 ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 20

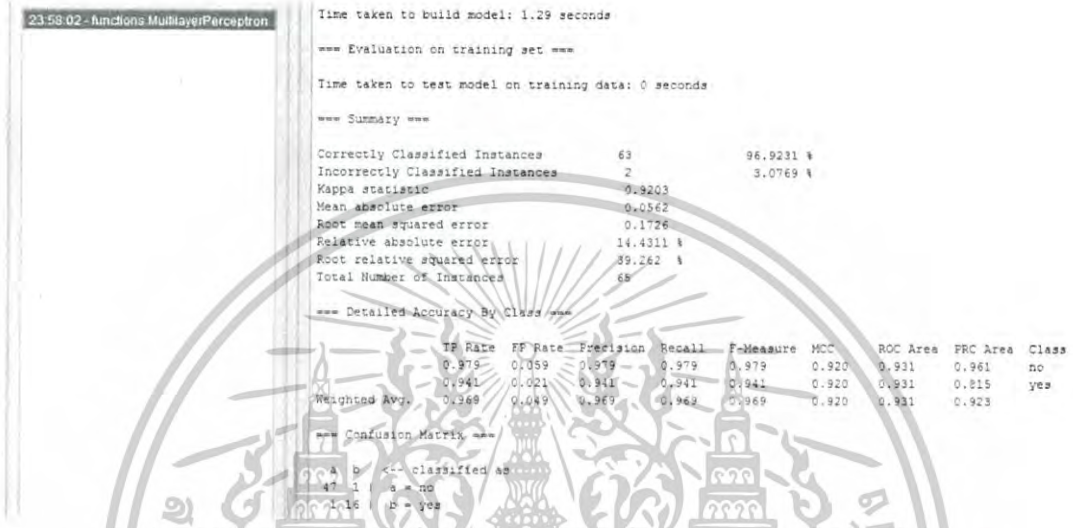
จากรูป 4-106 ส่มครั้งที่ 2 Random Seed = 20 กลุ่มลูกค้า 10 เปอร์เซ็นต์ จำนวนทั้งหมด 65 คน มีค่าทำนายความถูกต้องจำนวน 57 คน คิดเป็น 87.6923% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.3278)^2$ = 0.1075 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.2062



รูปที่ 4-107 โครงข่ายประสาท การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 30

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากรูป 4-107 โครงข่ายประสาท โรคมะเร็งเต้านมของรัฐวิสคอนซิน ชั้นข้อมูลเข้า อยู่ด้านซ้ายสุดซึ่งมีคุณลักษณะทั้งหมด 23 ตัวแปร (X_1-X_{23}) ชั้นซ่อน มี 1 ชั้นซ่อน ชั้นข้อมูลออกหรือชั้นข้อมูลผลลัพธ์ อยู่ด้านขวามือมีคำตอบคือ Yes กับ No Epoch จะรันการทำงานตั้งแต่รอบที่ 1 ถึง 20,000 Error per Epoch แสดงความคลาดเคลื่อนต่อรอบ คือ Learning Rate เท่ากับ 0.1 และ Momentum เท่ากับ 0.9



รูปที่ 4-108 ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 30

จากรูป 4-108 สุ่มครั้งที่ 3 Random Seed = 30 กลุ่มลูกค้า 10 เปอร์เซ็นต์ จำนวนทั้งหมด 65 คน มีค่าทำนายความถูกต้องจำนวน 63 คน คิดเป็น 96.9231% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.1726)^2$ = 0.0298 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0562

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.6 ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO

4.6.1 ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO โดยโปรแกรม SPSS

4.6.1.1 โรคมะเร็งเต้านมของรัฐวิสคอนซิน (Breast Cancer Wisconsin) จำนวนข้อมูลทดสอบ 699 ค่า ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่ต่ำ เมื่อกำหนดค่า Random Seed 10 20 และ 30 ตามลำดับ

23:51:55 - functions.SMO

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances          69               98.5714 %
Incorrectly Classified Instances         1                1.4286 %
Kappa statistic                          0.968
Mean absolute error                      0.0143
Root mean squared error                  0.1195
Relative absolute error                  3.2258 %
Root relative squared error             25.4457 %
Total Number of Instances               70

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.979     0.000     1.000     0.979     0.989     0.968     0.989     0.993     no
                 1.000     0.021     0.958     1.000     0.979     0.968     0.989     0.958     yes
Weighted Avg.    0.986     0.007     0.986     0.986     0.986     0.968     0.989     0.982

=== Confusion Matrix ===

  a  b   <-- classified as
 46  1 |  a = no
  0 23 |  b = yes
```

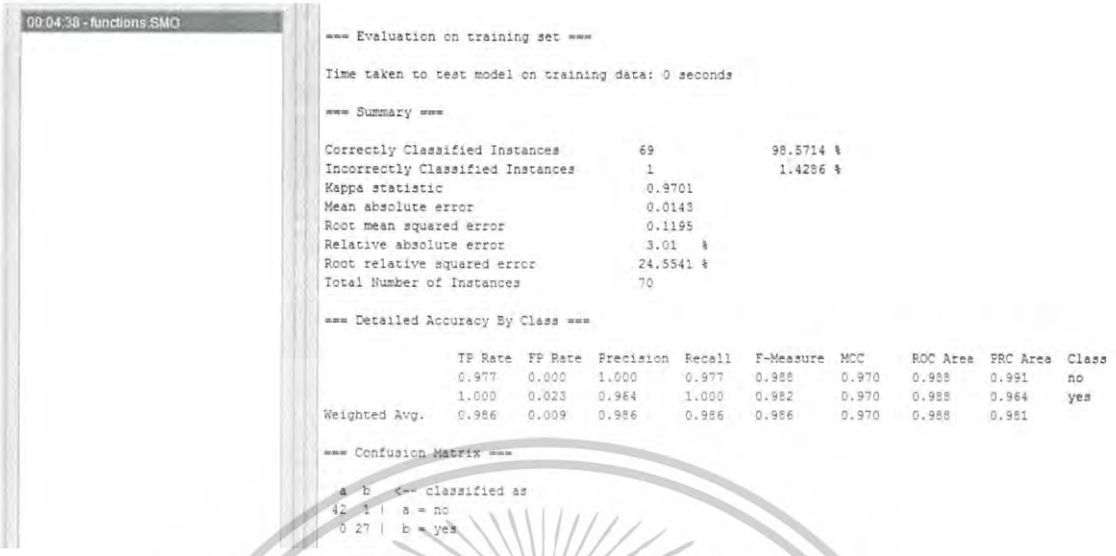
รูปที่ 4-109 ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 10

23:29:36 - functions.SMO

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances          69               98.5714 %
Incorrectly Classified Instances         1                1.4286 %
Kappa statistic                          0.968
Mean absolute error                      0.0143
Root mean squared error                  0.1195
Relative absolute error                  3.2258 %
Root relative squared error             25.4457 %
Total Number of Instances               70

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.979     0.000     1.000     0.979     0.989     0.968     0.989     0.993     no
                 1.000     0.021     0.958     1.000     0.979     0.968     0.989     0.958     yes
Weighted Avg.    0.986     0.007     0.986     0.986     0.986     0.968     0.989     0.982

=== Confusion Matrix ===

  a  b   <-- classified as
 46  1 |  a = no
  0 23 |  b = yes
```

รูปที่ 4-110 ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 20

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



00:04:38 - functions.SMO

```
=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          69               98.5714 %
Incorrectly Classified Instances         1                1.4286 %
Kappa statistic                          0.9701
Mean absolute error                      0.0143
Root mean squared error                  0.1195
Relative absolute error                  3.01   %
Root relative squared error             24.5541 %
Total Number of Instances               70

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.977    0.000    1.000      0.977   0.988      0.970    0.988     0.991     no
               1.000    0.023    0.964      1.000   0.982      0.970    0.988     0.964     yes
Weighted Avg.  0.986    0.009    0.986      0.986   0.986      0.970    0.988     0.981

=== Confusion Matrix ===

  a  b   <-- classified as
 42  1 |  a = no
  0 27 |  b = yes
```

รูปที่ 4-111 ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 30

จากรูป 4-109 ถึง 4-111 ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron โดยโปรแกรม SPSS โรคมะเร็งเต้านมของรัฐวิสคอนซิน โดยทำการสุ่ม 3 ครั้ง Random Seed 10 20 และ 30 ตามลำดับ จากผลลัพธ์ที่ได้พบว่า

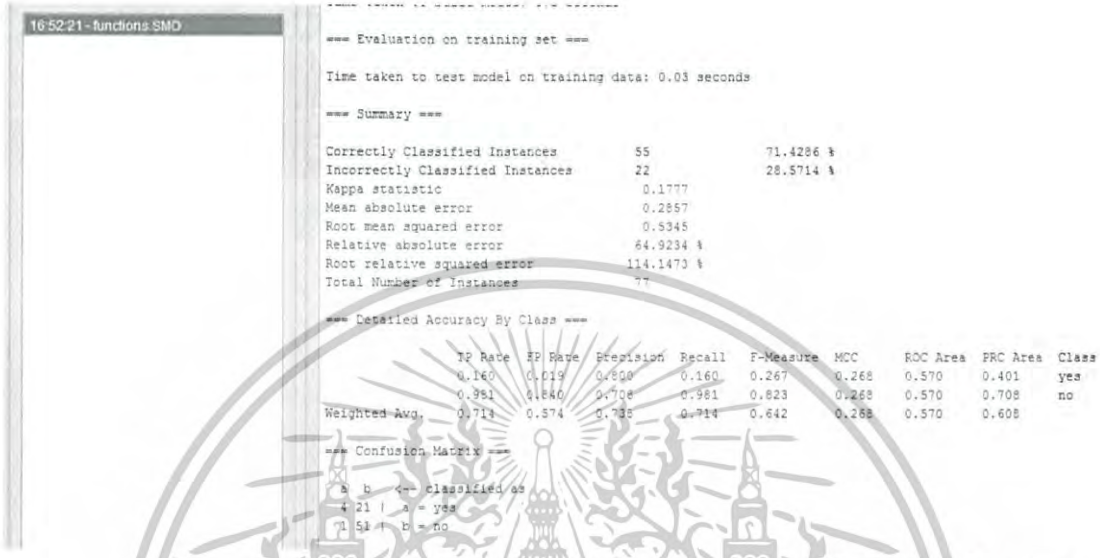
สุ่มครั้งที่ 1 Random Seed = 10 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 70 คน มีค่าทำนายความถูกต้องจำนวน 69 คน คิดเป็น 93.8462% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.1195)^2$ = 0.0566 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.1113

สุ่มครั้งที่ 2 Random Seed = 20 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 70 คน มีค่าทำนายความถูกต้องจำนวน 69 คน คิดเป็น 87.6923% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.1195)^2$ = 0.1075 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.2062

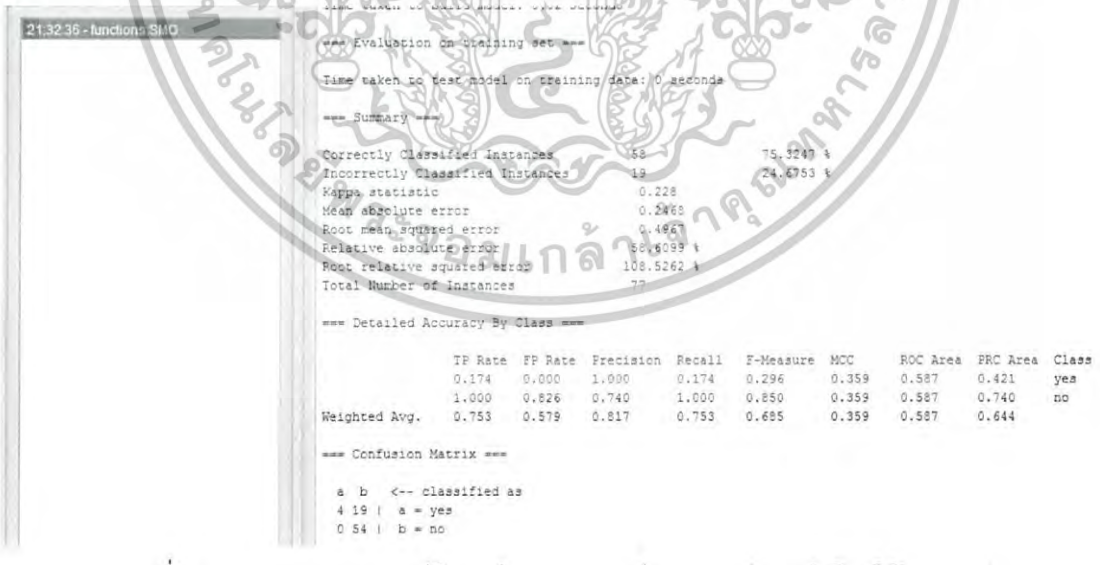
สุ่มครั้งที่ 3 Random Seed = 30 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 70 คน มีค่าทำนายความถูกต้องจำนวน 69 คน คิดเป็น 96.9231% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.1195)^2$ = 0.0298 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0562

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.6.1.2 โรคเบาหวานของชาวพิม่า ประเทศอินเดีย (Pima Indians Diabetes) จำนวนข้อมูลทดสอบ 768 ค่า ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่ปานกลาง เมื่อกำหนดค่า Random Seed 10 20 และ 30 ตามลำดับ

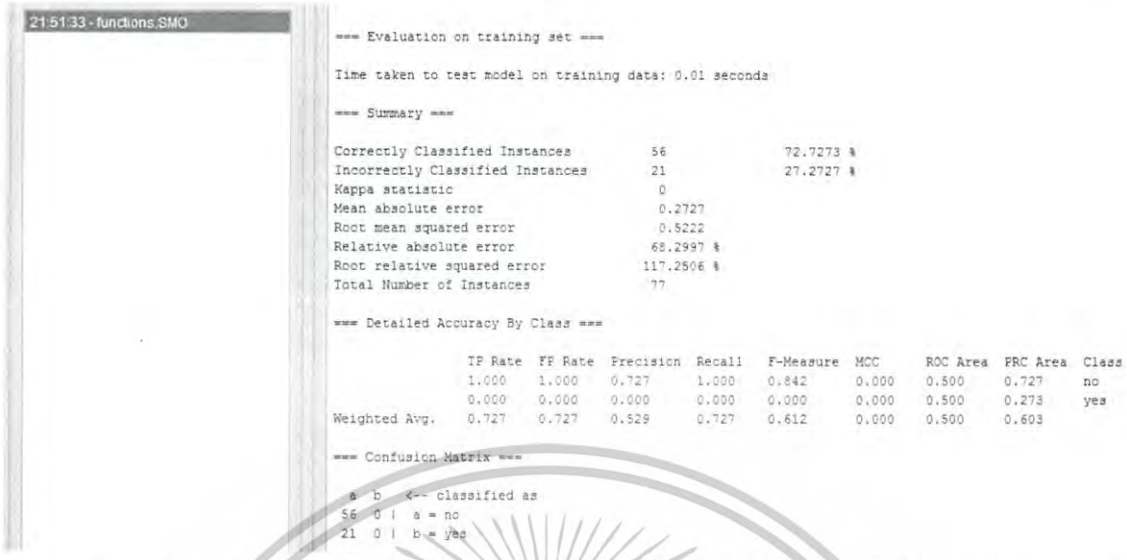


รูปที่ 4-112 ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 10



รูปที่ 4-113 ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 20

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



21:51:33 - functions.SMO

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances          56               72.7273 %
Incorrectly Classified Instances        21               27.2727 %
Kappa statistic                          0
Mean absolute error                      0.2727
Root mean squared error                  0.5222
Relative absolute error                 68.2997 %
Root relative squared error            117.2506 %
Total Number of Instances               77

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               1.000    1.000    0.727      1.000    0.842      0.000    0.500     0.727     no
               0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.273     yes
Weighted Avg.  0.727    0.727    0.529      0.727    0.612      0.000    0.500     0.603

=== Confusion Matrix ===

  a  b   <-- classified as
 56  0 |  a = no
 21  0 |  b = yes
```

รูปที่ 4-114 ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 30

จากรูป 4-112 ถึง 4-114 ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron โดยโปรแกรม SPSS โรคเบาหวานของชาวพิม่า ประเทศอินเดีย โดยทำการสุ่ม 3 ครั้ง Random Seed 10 20 และ 30 ตามลำดับ จากผลลัพธ์ที่ได้พบว่า

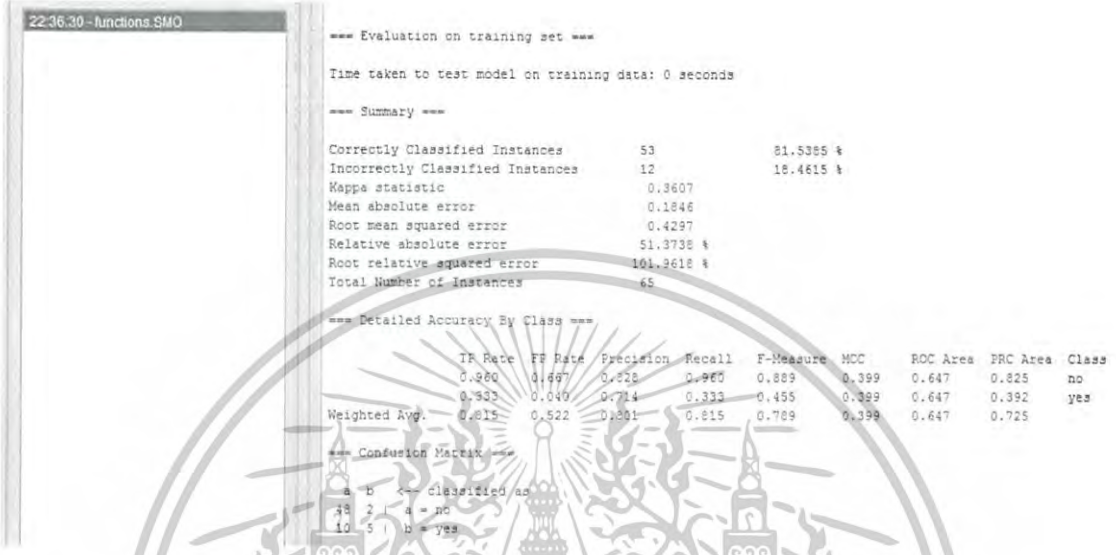
สุ่มครั้งที่ 1 Random Seed = 10 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 77 คน มีค่าทำนายความถูกต้องจำนวน 55 คน คิดเป็น 71.4286% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.5345)^2$ = 0.2857 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.2857

สุ่มครั้งที่ 2 Random Seed = 20 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 77 คน มีค่าทำนายความถูกต้องจำนวน 58 คน คิดเป็น 75.3247% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.4967)^2$ = 0.2467 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.2468

สุ่มครั้งที่ 3 Random Seed = 30 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 77 คน มีค่าทำนายความถูกต้องจำนวน 56 คน คิดเป็น 72.7273% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.5222)^2$ = 0.2727 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.2727

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

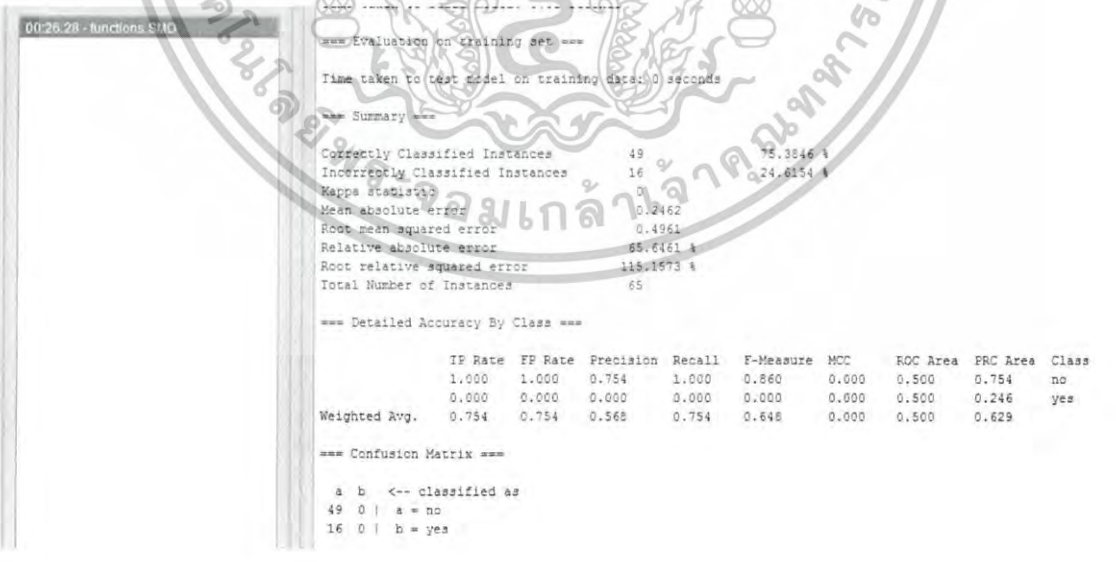
4.6.1.3 การชำระเงินด้วยบัตรเครดิตของลูกค้า (Default of Credit Card Clients) จำนวนข้อมูลทดสอบ 646 ค่า ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่สูง เมื่อกำหนดค่า Random Seed 10 20 และ 30 ตามลำดับ



22:36:30 - functions.SMO

```
=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          53               81.5385 %
Incorrectly Classified Instances        12               18.4615 %
Kappa statistic                          0.3607
Mean absolute error                      0.1846
Root mean squared error                  0.4297
Relative absolute error                 51.3738 %
Root relative squared error            101.9618 %
Total Number of Instances               65

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.960     0.667     0.828     0.960     0.889     0.399     0.647     0.825     no
                 0.333     0.040     0.714     0.333     0.455     0.399     0.647     0.392     yes
Weighted Avg.    0.815     0.522     0.801     0.815     0.789     0.399     0.647     0.725

=== Confusion Matrix ===

  a  b   <-- classified as
 48  2 |  a = no
 10  5 |  b = yes
```

รูปที่ 4-115 ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 10

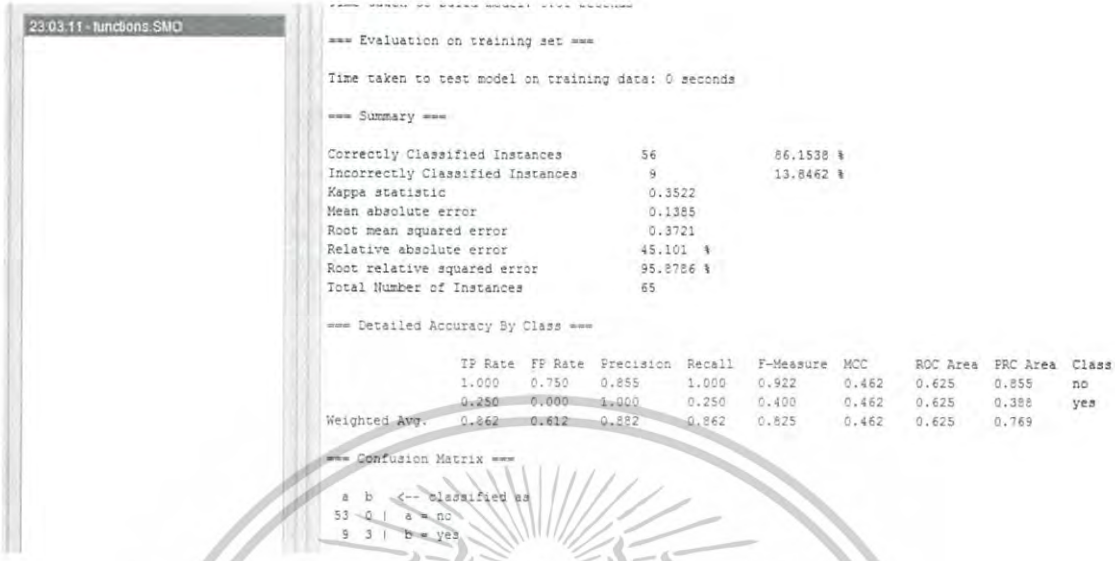


00:26:28 - functions.SMO

```
=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          49               75.3846 %
Incorrectly Classified Instances        16               24.6154 %
Kappa statistic                          0
Mean absolute error                      0.2462
Root mean squared error                  0.4961
Relative absolute error                 65.6461 %
Root relative squared error            115.1573 %
Total Number of Instances               65

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000     1.000     0.754     1.000     0.860     0.000     0.500     0.754     no
                 0.000     0.000     0.000     0.000     0.000     0.000     0.500     0.246     yes
Weighted Avg.    0.754     0.754     0.568     0.754     0.648     0.000     0.500     0.629

=== Confusion Matrix ===

  a  b   <-- classified as
 49  0 |  a = no
 16  0 |  b = yes
```

รูปที่ 4-116 ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 20

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



23:03:11 - functions.SMO

```
=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          56               86.1538 %
Incorrectly Classified Instances         9               13.8462 %
Kappa statistic                          0.3522
Mean absolute error                      0.1385
Root mean squared error                  0.3721
Relative absolute error                 45.101  %
Root relative squared error             95.8786 %
Total Number of Instances               65

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               1.000    0.750     0.855       1.000   0.922      0.462    0.625     0.855     no
               0.250    0.000     1.000       0.250   0.400      0.462    0.625     0.388     yes
Weighted Avg.  0.862    0.612     0.882       0.862   0.825      0.462    0.625     0.769

=== Confusion Matrix ===

  a  b   <-- classified as
 53  0 |  a = no
  9  3 |  b = yes
```

รูปที่ 4-117 ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 30

จากรูป 4-115 ถึง 4-117 ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron โดยโปรแกรม SPSS การชำระเงินด้วยบัตรเครดิตของลูกค้า โดยทำการสุ่ม 3 ครั้ง Random Seed 10 20 และ 30 ตามลำดับ จากผลลัพธ์ที่ได้พบว่า

สุ่มครั้งที่ 1 Random Seed = 10 กลุ่มลูกค้า 10 เปอร์เซ็นต์ จำนวนทั้งหมด 65 คน มีค่าทำนายความถูกต้องจำนวน 53 คน คิดเป็น 81.5385% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.4297)^2$ = 0.1846 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.1846

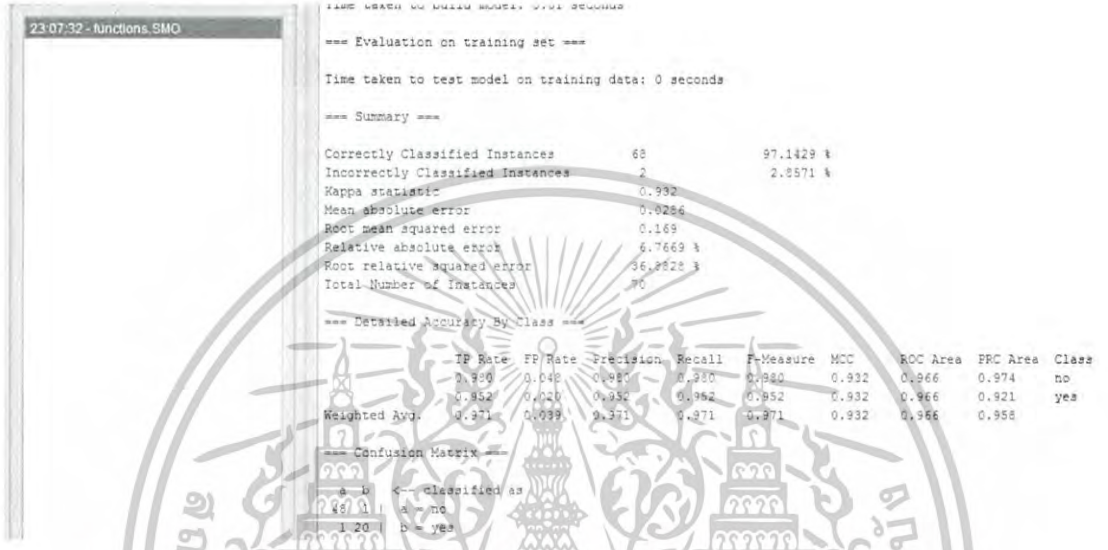
สุ่มครั้งที่ 2 Random Seed = 20 กลุ่มลูกค้า 10 เปอร์เซ็นต์ จำนวนทั้งหมด 65 คน มีค่าทำนายความถูกต้องจำนวน 49 คน คิดเป็น 75.3846% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.4961)^2$ = 0.2461 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.2462

สุ่มครั้งที่ 3 Random Seed = 30 กลุ่มลูกค้า 10 เปอร์เซ็นต์ จำนวนทั้งหมด 65 คน มีค่าทำนายความถูกต้องจำนวน 56 คน คิดเป็น 86.1538% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.3721)^2$ = 0.1385 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.1385

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.6.2 ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO โดยโปรแกรม WEKA

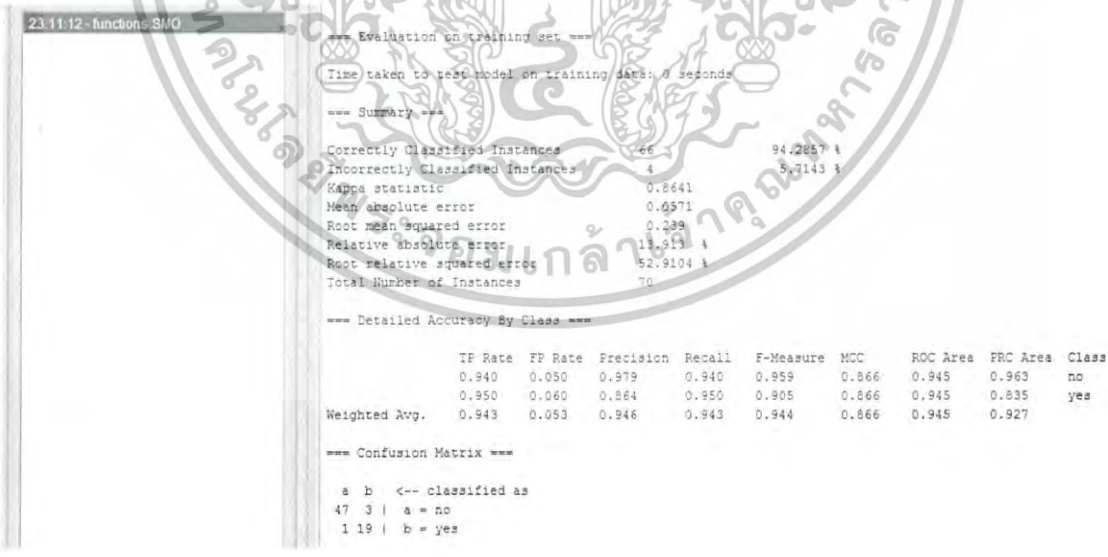
4.6.2.1 โรคมะเร็งเต้านมของรัฐวิสคอนซิน (Breast Cancer Wisconsin) จำนวนข้อมูลทดสอบ 699 ค่า ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่ต่ำ เมื่อกำหนดค่า Random Seed 10 20 และ 30 ตามลำดับ



23:07:32 - functions.SMO

```
=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          68               97.1429 %
Incorrectly Classified Instances         2                2.8571 %
Kappa statistic                          0.932
Mean absolute error                      0.0286
Root mean squared error                  0.169
Relative absolute error                  6.7669 %
Root relative squared error             36.8828 %
Total Number of Instances               70

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0.980     0.048      0.980     0.980     0.980     0.932     0.966     0.974     no
                 0.952     0.020      0.952     0.952     0.952     0.932     0.966     0.921     yes
Weighted Avg.    0.971     0.039      0.971     0.971     0.971     0.932     0.966     0.958

=== Confusion Matrix ===

  a  b   <-- classified as
 48  1 |  a = no
  1 20 |  b = yes
```

รูปที่ 4-118 ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 10

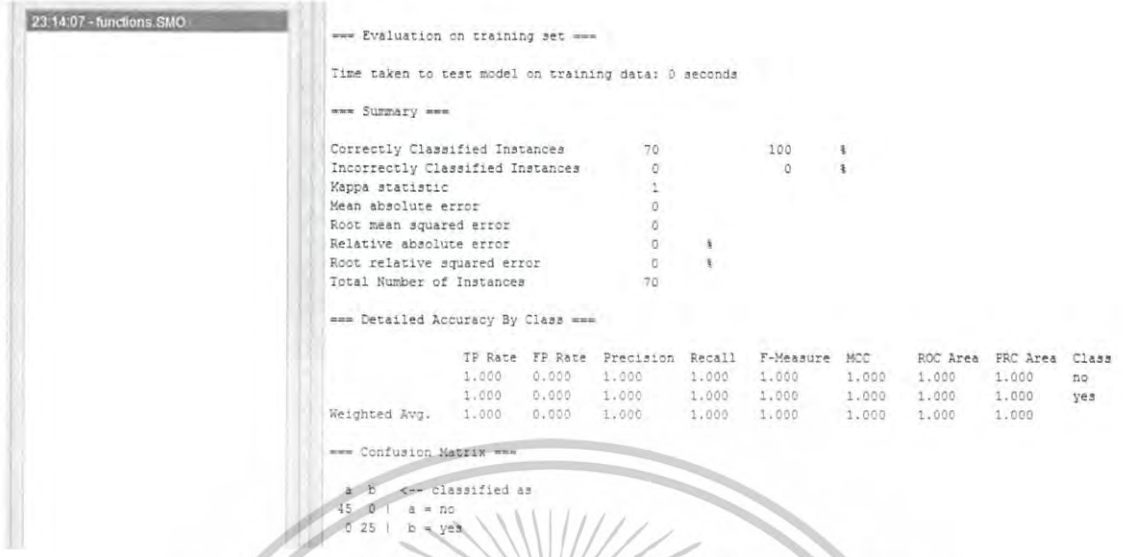


23:11:12 - functions.SMO

```
=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          66               94.2857 %
Incorrectly Classified Instances         4                5.7143 %
Kappa statistic                          0.8641
Mean absolute error                      0.0571
Root mean squared error                  0.239
Relative absolute error                 13.513  %
Root relative squared error             52.9104 %
Total Number of Instances               70

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0.940     0.050      0.979     0.940     0.959     0.866     0.945     0.963     no
                 0.950     0.060      0.864     0.950     0.905     0.866     0.945     0.835     yes
Weighted Avg.    0.943     0.053      0.946     0.943     0.944     0.866     0.945     0.927

=== Confusion Matrix ===

  a  b   <-- classified as
 47  3 |  a = no
  1 19 |  b = yes
```

รูปที่ 4-119 ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 20

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



23:14:07 - functions.SMO

```
=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          70               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances               70

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000   1.000      1.000    1.000     1.000     no
                 1.000    0.000    1.000      1.000   1.000      1.000    1.000     1.000     yes
Weighted Avg.    1.000    0.000    1.000      1.000   1.000      1.000    1.000     1.000

=== Confusion Matrix ===

  a  b   <-- classified as
 45  0 |  a = no
  0 25 |  b = yes
```

รูปที่ 4-120 ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO โรคมะเร็งเต้านมของรัฐวิสคอนซิน เมื่อ Random Seed 30

จากรูป 4-118 ถึง 4-120 ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron โดยโปรแกรม WEKA โรคมะเร็งเต้านมของรัฐวิสคอนซิน โดยทำการสุ่ม 3 ครั้ง Random Seed 10 20 และ 30 ตามลำดับ จากผลลัพธ์ที่ได้พบว่า

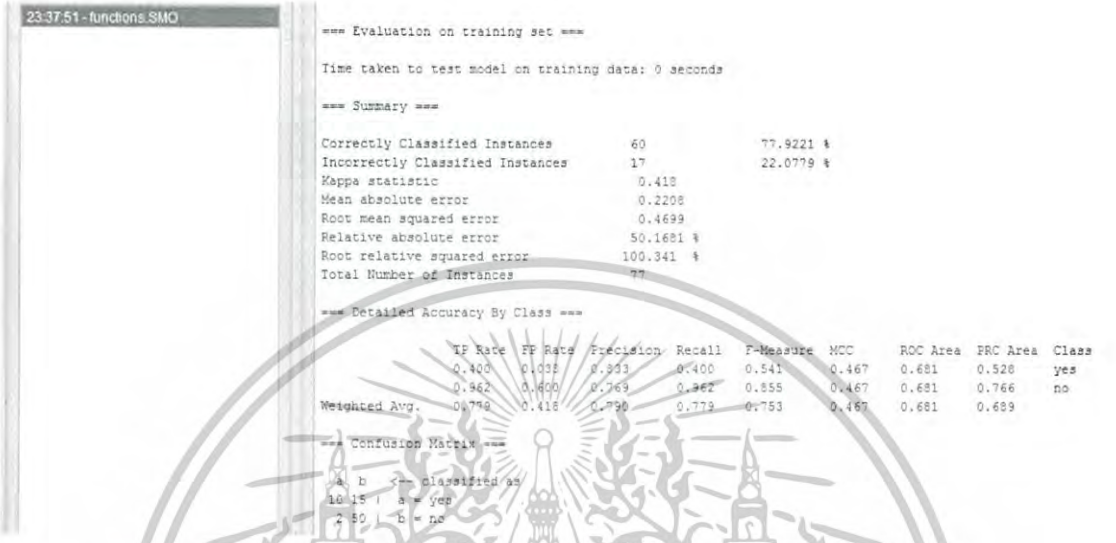
สุ่มครั้งที่ 1 Random Seed = 10 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 70 คน มีค่าทำนายความถูกต้องจำนวน 68 คน คิดเป็น 97.1429% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.169)^2$ = 0.0286 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0286

สุ่มครั้งที่ 2 Random Seed = 20 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 70 คน มีค่าทำนายความถูกต้องจำนวน 66 คน คิดเป็น 94.2857% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.239)^2$ = 0.0571 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.0571

สุ่มครั้งที่ 3 Random Seed = 30 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 70 คน มีค่าทำนายความถูกต้องจำนวน 70 คน คิดเป็น 100% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0)^2$ = 0 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0

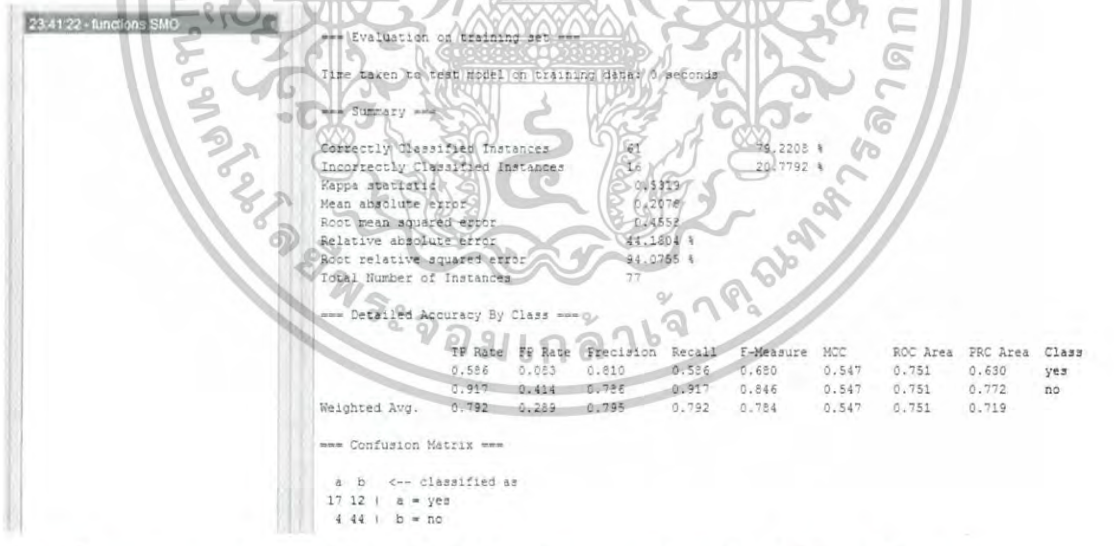
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.6.2.2 โรคเบาหวานของชาวพิม่า ประเทศอินเดีย (Pima Indians Diabetes) จำนวนข้อมูลทดสอบ 768 ค่า ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่ปานกลาง เมื่อกำหนดค่า Random Seed 10 20 และ 30 ตามลำดับ



```
23:37:51 - functions.SMO

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          60               77.9221 %
Incorrectly Classified Instances        17               22.0779 %
Kappa statistic                          0.418
Mean absolute error                      0.2208
Root mean squared error                  0.4699
Relative absolute error                 50.1681 %
Root relative squared error            100.341  %
Total Number of Instances               77

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.400    0.038    0.833      0.400   0.541      0.467    0.681     0.528     yes
                 0.962    0.600    0.769      0.962   0.855      0.467    0.681     0.766     no
Weighted Avg.    0.779    0.418    0.790      0.779   0.753      0.467    0.681     0.689

=== Confusion Matrix ===

  a  b   <-- classified as
 10 15 |  a = yes
  2 50 |  b = no
```

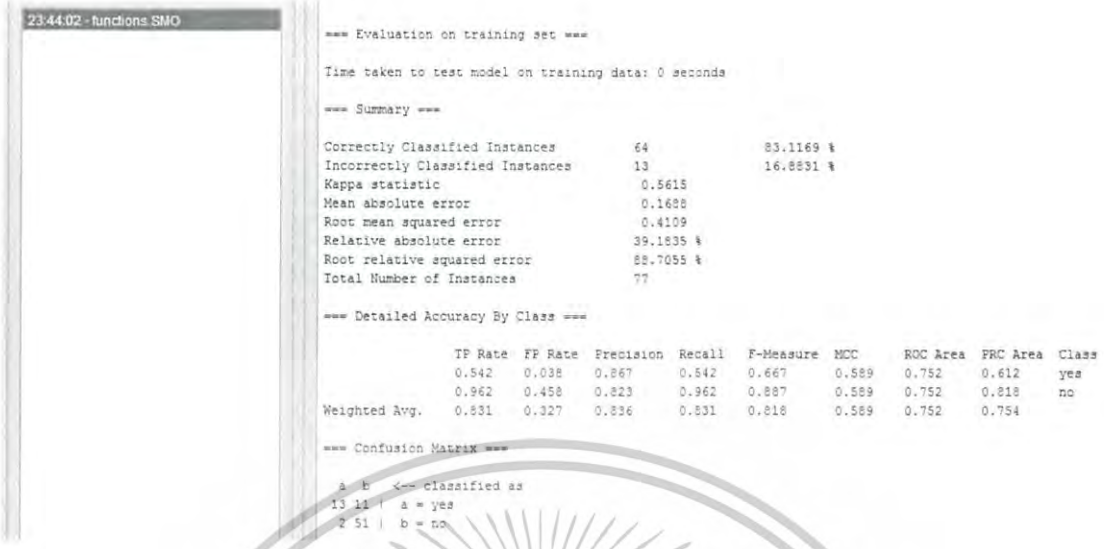
รูปที่ 4-121 ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 10



```
23:41:22 - functions.SMO

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          61               79.2208 %
Incorrectly Classified Instances        16               20.7792 %
Kappa statistic                          0.5313
Mean absolute error                      0.2076
Root mean squared error                  0.4552
Relative absolute error                 44.1804 %
Root relative squared error             94.0755 %
Total Number of Instances               77

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.586    0.083    0.810      0.586   0.680      0.547    0.751     0.630     yes
                 0.917    0.414    0.786      0.917   0.846      0.547    0.751     0.772     no
Weighted Avg.    0.792    0.289    0.795      0.792   0.784      0.547    0.751     0.719

=== Confusion Matrix ===

  a  b   <-- classified as
 17 12 |  a = yes
  4 44 |  b = no
```

รูปที่ 4-122 ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 20

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



23:44:02 - functions.SMO

```
=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          64               83.1169 %
Incorrectly Classified Instances        13               16.8831 %
Kappa statistic                          0.5615
Mean absolute error                      0.1688
Root mean squared error                  0.4109
Relative absolute error                 39.1835 %
Root relative squared error             88.7055 %
Total Number of Instances               77

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
               0.542     0.038     0.867       0.542   0.667       0.589    0.752     0.612     yes
               0.962     0.458     0.823       0.962   0.887       0.589    0.752     0.818     no
Weighted Avg.  0.831     0.327     0.836       0.831   0.818       0.589    0.752     0.754

=== Confusion Matrix ===

  a  b   <-- classified as
 13 11 |  a = yes
  2 51 |  b = no
```

รูปที่ 4-123 ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เมื่อ Random Seed 30

จากรูป 4-121 ถึง 4-123 ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron โดยโปรแกรม WEKA โรคเบาหวานของชาวพิม่า ประเทศอินเดีย โดยทำการสุ่ม 3 ครั้ง Random Seed 10 20 และ 30 ตามลำดับ จากผลลัพธ์ที่ได้พบว่า

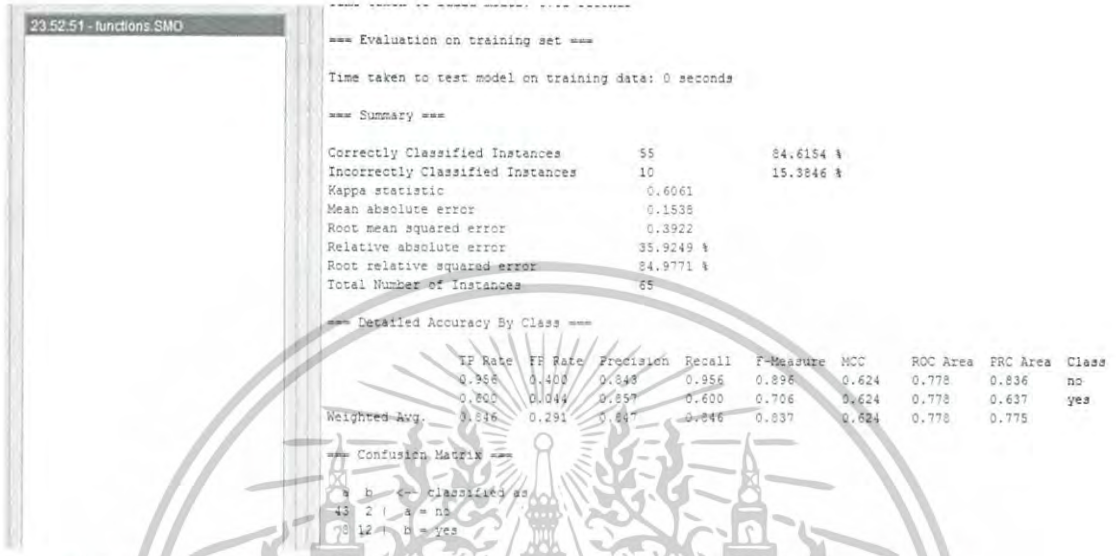
สุ่มครั้งที่ 1 Random Seed = 10 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 77 คน มีค่าทำนายความถูกต้องจำนวน 60 คน คิดเป็น 77.9221% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.4699)^2$ = 0.2208 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.2208

สุ่มครั้งที่ 2 Random Seed = 20 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 77 คน มีค่าทำนายความถูกต้องจำนวน 61 คน คิดเป็น 79.2208% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.4558)^2$ = 0.2078 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.2078

สุ่มครั้งที่ 3 Random Seed = 30 กลุ่มคนไข้ 10 เปอร์เซ็นต์ จำนวนทั้งหมด 77 คน มีค่าทำนายความถูกต้องจำนวน 64 คน คิดเป็น 83.1169% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.4109)^2$ = 0.1688 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.1688

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

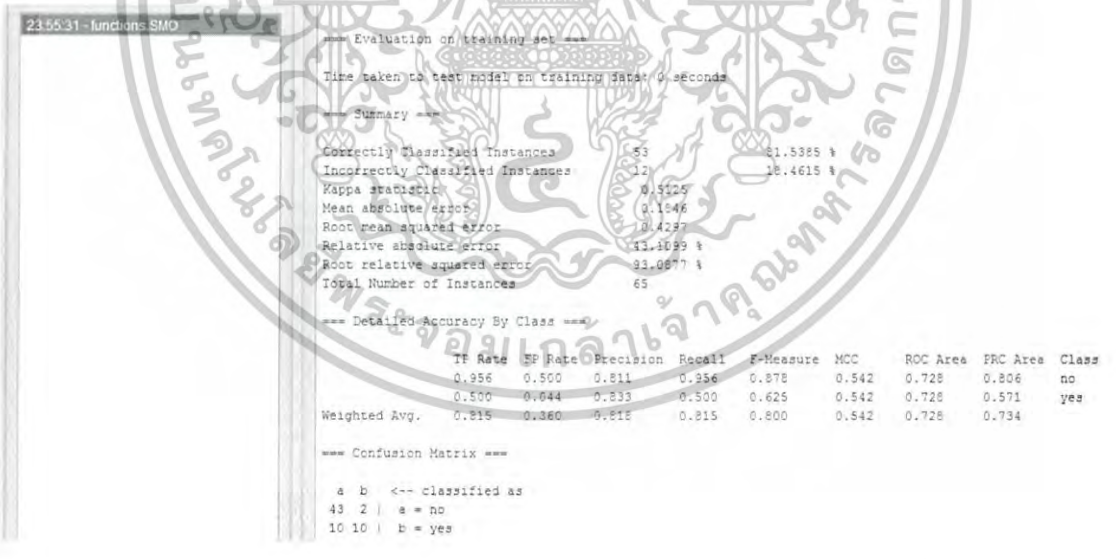
4.6.2.3 การชำระเงินด้วยบัตรเครดิตของลูกค้า (Default of Credit Card Clients) จำนวนข้อมูลทดสอบ 646 ค่า ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่สูง เมื่อกำหนดค่า Random Seed 10 20 และ 30 ตามลำดับ



23:52:51 - functions.SMO

```
=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          55               84.6154 %
Incorrectly Classified Instances        10               15.3846 %
Kappa statistic                          0.6061
Mean absolute error                      0.1538
Root mean squared error                  0.3922
Relative absolute error                 35.9249 %
Root relative squared error             84.9771 %
Total Number of Instances               65

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0.956     0.400     0.843     0.956     0.896     0.624     0.778     0.836     no
                 0.600     0.044     0.857     0.600     0.706     0.624     0.778     0.637     yes
Weighted Avg.    0.846     0.291     0.847     0.846     0.837     0.624     0.778     0.775

=== Confusion Matrix ===

  a  b   <-- classified as
 43  2 |  a = no
  8 12 |  b = yes
```

รูปที่ 4-124 ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 10

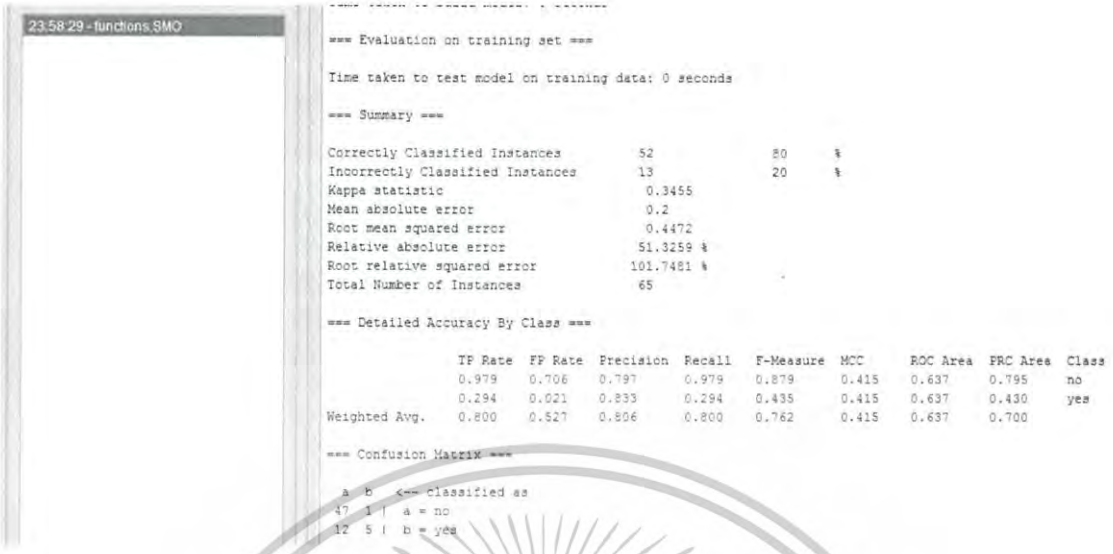


23:55:31 - functions.SMO

```
=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          53               81.5385 %
Incorrectly Classified Instances        12               18.4615 %
Kappa statistic                          0.5125
Mean absolute error                      0.1846
Root mean squared error                  0.4297
Relative absolute error                 43.1099 %
Root relative squared error             93.0877 %
Total Number of Instances               65

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0.956     0.500     0.811     0.956     0.878     0.542     0.728     0.806     no
                 0.500     0.044     0.833     0.500     0.625     0.542     0.728     0.571     yes
Weighted Avg.    0.815     0.360     0.818     0.815     0.800     0.542     0.728     0.734

=== Confusion Matrix ===

  a  b   <-- classified as
 43  2 |  a = no
 10 10 |  b = yes
```

รูปที่ 4-125 ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 20

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4-126 ผลการวิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO การชำระเงินด้วยบัตรเครดิตของลูกค้า เมื่อ Random Seed 30

จากรูป 4-124 ถึง 4-126 ผลการวิเคราะห์ข้อมูลด้วยวิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึม Multilayer Perceptron โดยโปรแกรม WEKA การชำระเงินด้วยบัตรเครดิตของลูกค้า โดยทำการสุ่ม 3 ครั้ง Random Seed 10 20 และ 30 ตามลำดับ จากผลลัพธ์ที่ได้พบว่า

สุ่มครั้งที่ 1 Random Seed = 10 กลุ่มลูกค้า 10 เปอร์เซ็นต์ จำนวนทั้งหมด 65 คน มีค่าทำนายความถูกต้องจำนวน 55 คน คิดเป็น 84.6154% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.3922)^2$ = 0.1538 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.1538

สุ่มครั้งที่ 2 Random Seed = 20 กลุ่มลูกค้า 10 เปอร์เซ็นต์ จำนวนทั้งหมด 65 คน มีค่าทำนายความถูกต้องจำนวน 53 คน คิดเป็น 81.5385% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.4297)^2$ = 0.1846 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.1846

สุ่มครั้งที่ 3 Random Seed = 30 กลุ่มลูกค้า 10 เปอร์เซ็นต์ จำนวนทั้งหมด 65 คน มีค่าทำนายความถูกต้องจำนวน 52 คน คิดเป็น 80% โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย = $(0.4472)^2$ = 0.1990 และมีค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย = 0.2

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.7 ผลการเปรียบเทียบประสิทธิภาพของวิธีการจำแนกกลุ่มโดยใช้โปรแกรม SPSS และ WEKA

4.7.1 ชุดข้อมูลโรคมะเร็งเต้านมของรัฐวิสคอนซิน ซึ่งเป็นชุดข้อมูลที่มีค่านอกเกณฑ์ระดับน้อย

ตารางที่ 4-4 ตารางชุดข้อมูลโรคมะเร็งเต้านมของรัฐวิสคอนซิน ซึ่งเป็นชุดข้อมูลที่มีค่านอกเกณฑ์ระดับน้อย โดยการสุ่มตัวอย่าง seed 10

วิธีการจำแนกกลุ่ม	โปรแกรม	ค่าความถูกต้อง	ความคลาดเคลื่อนกำลังสองเฉลี่ย	ส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย
นาอีฟเบส์	SPSS	97.1429	0.0281	0.0284
	WEKA	97.1429	0.0282	0.0285
เพื่อนบ้านใกล้สุด k ตัว	SPSS	100	0.0002	0.0121
	WEKA	100	0.0005	0.0115
ต้นไม้ตัดสินใจ	SPSS	98.5714	0.0114	0.0229
	WEKA	98.5714	0.0135	0.0271
โครงข่ายประสาทเทียม	SPSS	100	0.0000	0.0013
	WEKA	98.5714	0.0136	0.0274
ซัพพอร์ตเวกเตอร์แมชชีน	SPSS	98.5714	0.0143	0.0143
	WEKA	97.1429	0.0286	0.0286

จากตารางที่ 4-4 วิธีเพื่อนบ้านใกล้สุด k ตัว การสุ่มด้วยโปรแกรม SPSS วิธีต้นไม้ตัดสินใจ การสุ่มด้วยโปรแกรม WEKA วิธีโครงข่างประสาทเทียม การสุ่มด้วยโปรแกรม SPSS ให้ค่าความถูกต้องสูงสุด คือ 100% วิธีโครงข่ายประสาทเทียม การสุ่มด้วยโปรแกรม SPSS ให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.0000 และ ให้ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ยต่ำสุด คือ 0.0013

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4-5 ตารางชุดข้อมูลโรคมะเร็งเต้านมของรัฐวิสคอนซิน ซึ่งเป็นชุดข้อมูลที่มีค่านอกเกณฑ์ระดับน้อย โดยการสุ่มตัวอย่าง seed 20

วิธีการจำแนกกลุ่ม	โปรแกรม	ค่าความถูกต้อง	ความคลาดเคลื่อนกำลังสองเฉลี่ย	ส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย
นาอีฟเบส์	SPSS	95.7143	0.0427	0.0428
	WEKA	94.2857	0.0608	0.0665
เพื่อนบ้านใกล้สุด k ตัว	SPSS	100	0.0002	0.0117
	WEKA	100	0.00016	0.0121
ต้นไม้ตัดสินใจ	SPSS	98.5714	0.0107	0.0214
	WEKA	94.2857	0.0510	0.1020
โครงข่ายประสาทเทียม	SPSS	100	0.0000	0.0013
	WEKA	98.5714	0.0136	0.0272
ซัพพอร์ตเวกเตอร์แมชชีน	SPSS	98.5714	0.0143	0.0143
	WEKA	94.2857	0.0571	0.0571

จากตาราง 4-5 วิธีเพื่อนบ้านใกล้สุด k ตัว การสุ่มด้วยโปรแกรม SPSS และ WEKA และวิธีโครงข่ายประสาทเทียม การสุ่มด้วยโปรแกรม SPSS ให้ค่าความถูกต้องสูงสุดคือ 100% วิธีโครงข่ายประสาทเทียม การสุ่มด้วยโปรแกรม SPSS ให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด 0.0000 และให้ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ยต่ำสุดคือ 0.0013

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4-6 ตารางชุดข้อมูลโรคมะเร็งเต้านมของรัฐวิสคอนซิน ซึ่งเป็นชุดข้อมูลที่มีค่านอกเกณฑ์ระดับน้อย โดยการสุ่มตัวอย่าง seed 30

วิธีการจำแนกกลุ่ม	โปรแกรม	ค่าความถูกต้อง	ความคลาดเคลื่อนกำลังสองเฉลี่ย	ส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย
นาอีฟเบส์	SPSS	95.7143	0.0430	0.0465
	WEKA	98.5714	0.0143	0.0143
เพื่อนบ้านใกล้สุด k ตัว	SPSS	100	0.0002	0.0133
	WEKA	100	0.0002	0.0119
ต้นไม้ตัดสินใจ	SPSS	98.5714	0.0139	0.0279
	WEKA	100	0	0
โครงข่ายประสาทเทียม	SPSS	100	0.0000	0.0013
	WEKA	100	0.0000	0.0013
ซัพพอร์ตเวกเตอร์แมชชีน	SPSS	98.5714	0.0143	0.0143
	WEKA	100	0	0

จากตาราง 4-6 วิธีเพื่อนบ้านใกล้สุด k ตัว การสุ่มด้วยโปรแกรม SPSS และ WEKA วิธีต้นไม้ตัดสินใจ การสุ่มด้วยโปรแกรม WEKA วิธีโครงข่ายประสาทเทียม การสุ่มด้วยโปรแกรม SPSS และ WEKA วิธีซัพพอร์ทเวกเตอร์แมชชีน การสุ่มด้วยโปรแกรม WEKA ให้ค่าความถูกต้องสูงสุด คือ 100% วิธีต้นไม้ตัดสินใจและวิธีซัพพอร์ทเวกเตอร์แมชชีน การสุ่มด้วยโปรแกรม WEKA ให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุดคือ 0 และให้ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ยต่ำสุดคือ 0

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.7.2 ชุดข้อมูลโรคเบาหวานของชาวพิม่า ประเทศอินเดีย ซึ่งเป็นชุดข้อมูลที่มีค่านอกเกณฑ์ระดับปานกลาง

ตารางที่ 4-7 ตารางชุดข้อมูลโรคเบาหวานของชาวพิม่า ประเทศอินเดีย ซึ่งเป็นชุดข้อมูลที่มีค่านอกเกณฑ์ระดับปานกลาง โดยการสุ่มตัวอย่าง seed 10

วิธีการจำแนกกลุ่ม	โปรแกรม	ค่าความถูกต้อง	ความคลาดเคลื่อนกำลังสองเฉลี่ย	ส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย
นาอีฟเบส์	SPSS	75.3247	0.1648	0.2930
	WEKA	79.2208	0.1638	0.2409
เพื่อนบ้านใกล้สุด k ตัว	SPSS	100	0.0002	0.0127
	WEKA	100	0.0002	0.0127
ต้นไม้ตัดสินใจ	SPSS	87.0130	0.1041	0.2081
	WEKA	81.8182	0.1482	0.2964
โครงข่ายประสาทเทียม	SPSS	87.0130	0.1176	0.2227
	WEKA	88.3117	0.1002	0.1955
ซัพพอร์ตเวกเตอร์แมชชีน	SPSS	71.4286	0.2857	0.2857
	WEKA	77.9221	0.2208	0.2208

จากตารางที่ 4-7 วิธีเพื่อนบ้านใกล้สุด k ตัว การสุ่มด้วยโปรแกรม SPSS และ WEKA ให้ค่าความถูกต้องสูงสุดคือ100% ให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุดคือ 0.0002 ให้ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ยต่ำสุดคือ 0.0127

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4-8 ตารางชุดข้อมูลโรคเบาหวานของชาวพิม่า ประเทศอินเดีย ซึ่งเป็นชุดข้อมูลที่มีค่านอกเกณฑ์ระดับปานกลาง โดยการสุ่มตัวอย่าง seed 20

วิธีการจำแนกกลุ่ม	โปรแกรม	ค่าความถูกต้อง	ความคลาดเคลื่อนกำลังสองเฉลี่ย	ส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย
นาอีฟเบส์	SPSS	72.7273	0.2031	0.2923
	WEKA	76.6234	0.1736	0.252
เพื่อนบ้านใกล้สุด k ตัว	SPSS	100	0.0002	0.0127
	WEKA	100	0.0002	0.0127
ต้นไม้ตัดสินใจ	SPSS	84.4156	0.1257	0.2514
	WEKA	88.3117	0.0892	0.1784
โครงข่ายประสาทเทียม	SPSS	84.4156	0.1321	0.2554
	WEKA	89.6104	0.0920	0.1751
ซัพพอร์ตเวกเตอร์แมชชีน	SPSS	75.3247	0.2467	0.2468
	WEKA	79.2208	0.2077	0.2078

จากตาราง 4-8 วิธีเพื่อนบ้านใกล้สุด k ตัว การสุ่มด้วยโปรแกรม SPSS และ WEKA ให้ค่าความถูกต้องสูงสุด คือ 100% ให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุดคือ 0.0002 และให้ค่าส่วนเบี่ยงเบนกำลังสองเฉลี่ยต่ำสุดคือ 0.0127

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4-9 ตารางชุดข้อมูลโรคเบาหวานของชาวพิม่า ประเทศอินเดีย ซึ่งเป็นชุดข้อมูลที่มีค่านอกเกณฑ์ระดับปานกลาง โดยการสุ่มตัวอย่าง seed 30

วิธีการจำแนกกลุ่ม	โปรแกรม	ค่าความถูกต้อง	ความคลาดเคลื่อนกำลังสองเฉลี่ย	ส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย
นาอีฟเบส์	SPSS	74.0260	0.1722	0.2723
	WEKA	81.8182	0.1443	0.2109
เพื่อนบ้านใกล้สุด k ตัว	SPSS	100	0.0002	0.0127
	WEKA	100	0.0002	0.0127
ต้นไม้ตัดสินใจ	SPSS	93.5065	0.0556	0.1112
	WEKA	87.0130	0.1092	0.2185
โครงข่ายประสาทเทียม	SPSS	85.7143	0.1220	0.2456
	WEKA	92.2078	0.0701	0.1387
ซัพพอร์ตเวกเตอร์แมชชีน	SPSS	72.7273	0.2727	0.2727
	WEKA	83.1169	0.1688	0.1688

จากตาราง 4-9 วิธีเพื่อนบ้านใกล้สุด k ตัว การสุ่มด้วยโปรแกรม SPSS และ WEKA ให้ค่าความถูกต้องสูงสุดคือ 100% ให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุดคือ 0.0002 และให้ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ยต่ำสุดคือ 0.0127

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.7.3 ชุดข้อมูลการชำระเงินด้วยบัตรเครดิตของลูกค้า ซึ่งเป็นข้อมูลที่มีค่านอกเกณฑ์ระดับมาก

ตารางที่ 4-10 ตารางชุดข้อมูลการชำระเงินด้วยบัตรเครดิตของลูกค้า ซึ่งเป็นข้อมูลที่มีค่านอกเกณฑ์ระดับมาก โดยการสุ่มตัวอย่าง seed 10

วิธีการจำแนกกลุ่ม	โปรแกรม	ค่าความถูกต้อง	ความคลาดเคลื่อนกำลังสองเฉลี่ย	ส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย
นาอีฟเบส์	SPSS	53.8462	0.4444	0.4579
	WEKA	67.6923	0.3085	0.3329
เพื่อนบ้านใกล้สุด k ตัว	SPSS	100	0.0002	0.0149
	WEKA	100	0.0002	0.0149
ต้นไม้ตัดสินใจ	SPSS	96.9231	0.0284	0.0568
	WEKA	95.3846	0.0393	0.0785
โครงข่ายประสาทเทียม	SPSS	92.3077	0.0700	0.1381
	WEKA	93.8462	0.0565	0.1113
ซัพพอร์ตเวกเตอร์แมชชีน	SPSS	81.5385	0.1846	0.1846
	WEKA	84.6154	0.1538	0.1538

จากตาราง 4-10 วิธีเพื่อนบ้านใกล้สุด k ตัว การสุ่มด้วยโปรแกรม SPSS และ WEKA ให้ค่าความถูกต้องสูงสุดคือ 100% ให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุดคือ 0.0002 และให้ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ยต่ำสุดคือ 0.0149

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4-11 ตารางชุดข้อมูลการชำระเงินด้วยบัตรเครดิตของลูกค้า ซึ่งเป็นข้อมูลที่มีค่านอกเกณฑ์ระดับมาก โดยการสุ่มตัวอย่าง seed 20

วิธีการจำแนกกลุ่ม	โปรแกรม	ค่าความถูกต้อง	ความคลาดเคลื่อนกำลังสองเฉลี่ย	ส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย
นาอีฟเบส์	SPSS	53.8462	0.4331	0.4491
	WEKA	72.3077	0.2236	0.2790
เพื่อนบ้านใกล้สุด k ตัว	SPSS	100	0.0002	0.0149
	WEKA	100	0.0002	0.0149
ต้นไม้ตัดสินใจ	SPSS	96.9231	0.0286	0.0571
	WEKA	89.2308	0.0900	0.1800
โครงข่ายประสาทเทียม	SPSS	87.6923	0.0823	0.1670
	WEKA	87.6923	0.1075	0.2062
ซัพพอร์ตเวกเตอร์แมชชีน	SPSS	75.3846	0.2461	0.2462
	WEKA	81.5385	0.1846	0.1846

จากตาราง 4-11 วิธีเพื่อนบ้านใกล้สุด k ตัว การสุ่มด้วยโปรแกรม SPSS และ WEKA ให้ค่าความถูกต้องสูงสุดคือ 100% ให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด 0.0002 และให้ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ยต่ำสุดคือ 0.0149

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4-12 ตารางชุดข้อมูลการชำระเงินด้วยบัตรเครดิตของลูกค้า ซึ่งเป็นข้อมูลที่มีค่านอกเกณฑ์ระดับมาก โดยการสุ่มตัวอย่าง seed 30

วิธีการจำแนกกลุ่ม	โปรแกรม	ค่าความถูกต้อง	ความคลาดเคลื่อนกำลังสอง	ส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย
นาอีฟเบส์	SPSS	60	0.3714	0.4011
	WEKA	61.5385	0.3049	0.3465
เพื่อนบ้านใกล้สุด k ตัว	SPSS	100	0.0002	0.0149
	WEKA	100	0.0002	0.0149
ต้นไม้ตัดสินใจ	SPSS	89.2308	0.0960	0.1919
	WEKA	98.4615	0.0103	0.0205
โครงข่ายประสาทเทียม	SPSS	95.3846	0.0438	0.0857
	WEKA	96.9231	0.0298	0.0562
ซัพพอร์ตเวกเตอร์แมชชีน	SPSS	86.1538	0.1385	0.1385
	WEKA	80	0.1999	0.2000

จากตาราง 4-12 วิธีเพื่อนบ้านใกล้สุด k ตัว การสุ่มด้วยโปรแกรม SPSS และ WEKA ให้ค่าความถูกต้องสูงสุด 100% ให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุดคือ 0.0002 และให้ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ยต่ำสุด 0.0149

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 5
สรุปผล อภิปรายผล และข้อเสนอแนะ

5.1 สรุปผลการวิจัย

ในการจัดทำปัญหาพิเศษครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพในการจำแนกกลุ่ม 5 วิธี คือ วิธีนาอีฟเบส์ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม และวิธีซัพพอร์ตเวกเตอร์แมชชีน รวมทั้งเปรียบเทียบวิธีการสุ่มตัวอย่างโดยใช้โปรแกรม SPSS กับ WEKA ว่าโปรแกรมใดให้ค่าความถูกต้องสูงกว่า ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยและค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ยต่ำกว่า เนื่องจากวิธีดังกล่าวข้างต้นเป็นวิธีที่นิยมนำมาใช้ในการจำแนกกลุ่มข้อมูลที่ทราบผลลัพธ์แน่นอน โดยใช้หลักการของการทำเหมืองข้อมูลมาใช้ในการจำแนกกลุ่มข้อมูล

จากการค้นคว้าและศึกษาในการหาค่านอกเกณฑ์ได้ข้อมูล 3 ชุด คือโรคมะเร็งเต้านมของรัฐวิสคอนซิน (Breast Cancer Wisconsin) เป็นชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับต่ำ โรคเบาหวานของชาวพิม่า ประเทศอินเดีย (Pima Indians Diabetes) เป็นชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับปานกลาง และการชำระเงินด้วยบัตรเครดิตของลูกค้า (Default Of Credit Card Clients) เป็นชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับสูง

โรคมะเร็งเต้านมของรัฐวิสคอนซิน Random Seed 10, 20 วิธีที่มีประสิทธิภาพสูงสุดคือ วิธีโครงข่ายประสาทเทียมของโปรแกรม SPSS มีค่าความถูกต้องคือ 100% ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยคือ 0 ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ยคือ 0.0013 และ Random Seed 30 วิธีที่มีประสิทธิภาพสูงสุดคือ วิธีต้นไม้ตัดสินใจและวิธีซัพพอร์ตเวกเตอร์แมชชีนของโปรแกรม WEKA มีค่าความถูกต้องคือ 100% ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยคือ 0 ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ยคือ 0 โรคเบาหวานของชาวพิม่า ประเทศอินเดีย Random Seed 10, 20, 30 วิธีที่มีประสิทธิภาพสูงสุดคือ วิธีเพื่อนบ้านใกล้สุด k ตัวของโปรแกรม SPSS และ WEKA มีค่าความถูกต้องคือ 100% ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยคือ 0.0002 ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ยคือ 0.0127 และการชำระเงินด้วยบัตรเครดิตของลูกค้า Random Seed 10, 20, 30 วิธีที่มีประสิทธิภาพสูงสุดคือ วิธีเพื่อนบ้านใกล้สุด k ตัวของโปรแกรม SPSS และ WEKA มีค่าความถูกต้องคือ 100% ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยคือ 0.0002 ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ยคือ 0.0149

ดังนั้นโรคมะเร็งเต้านมของรัฐวิสคอนซิน วิธีที่มีประสิทธิภาพสูงสุดคือ วิธีโครงข่ายประสาทเทียม โดยการสุ่มของโปรแกรม SPSS โรคเบาหวานของชาวพิม่า ประเทศอินเดีย วิธีที่มีประสิทธิภาพสูงสุดคือ วิธีเพื่อนบ้านใกล้สุด k ตัว โดยการสุ่มของโปรแกรม SPSS และ WEKA และการชำระเงินด้วยบัตรเครดิตของลูกค้า วิธีที่มีประสิทธิภาพสูงสุดคือ วิธีเพื่อนบ้านใกล้สุด k ตัว โดยการสุ่มของโปรแกรม SPSS และ WEKA ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับปานกลางและสูงให้ผลการจำแนกกลุ่มที่เหมือนกัน ซึ่งแตกต่างจากชุดข้อมูลที่มีค่านอกเกณฑ์ในระดับที่ต่ำ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

5.2 อภิปรายผล

จากการสรุปผลปัญหาพิเศษครั้งนี้ ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับต่ำ โรคมะเร็งเต้านมของรัฐวิสคอนซิน วิธีโครงข่ายประสาทเทียมใช้อัลกอริทึมชนิดเพอร์เซปตรอนแบบหลายชั้น โดยกำหนดค่าอัตราการเรียนรู้เป็น 0.1 ค่าโมเมนตัมเป็น 0.9 จำนวนรอบการสอน 20,000 รอบ และชั้นซ่อน 1 ชั้น มีประสิทธิภาพในการจำแนกกลุ่มดีที่สุด เพราะว่ามีค่าความถูกต้องสูงที่สุด ความคลาดเคลื่อนกำลังสองเฉลี่ย และค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ยต่ำที่สุด ให้ผลสอดคล้องกับ กิตติพล วิแสง และคณะ (2552) เรื่องการวิเคราะห์ปัจจัยเสี่ยงของโรคเบาหวาน (ชุดข้อมูลนี้ไม่ได้พิจารณาค่านอกเกณฑ์) วิธีโครงข่ายประสาทเทียม ให้ผลลัพธ์ค่าความถูกต้องสูงที่สุด ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับปานกลาง โรคเบาหวานของชาวพิม่า ประเทศอินเดียและชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับสูง การชำระเงินด้วยบัตรเครดิตของลูกค้า วิธีเพื่อนบ้านใกล้สุด k ตัว ใช้อัลกอริทึมชนิด IBK มีประสิทธิภาพในการจำแนกกลุ่มดีที่สุด เพราะว่ามีค่าความถูกต้องสูงที่สุด ความคลาดเคลื่อนกำลังสองเฉลี่ย และค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ยต่ำที่สุด ให้ผลสอดคล้องกับ นิเวศ จิระวิชิตชัย (2553) เรื่องการค้นหาเทคนิคเหมืองข้อมูลเพื่อสร้างโมเดลการวิเคราะห์โรคอัตโนมัติ โรคเบาหวาน (ชุดข้อมูลนี้ไม่ได้พิจารณาค่านอกเกณฑ์) ในกลุ่มข้อมูล Hepatitis วิธีเพื่อนบ้านใกล้สุด k ตัว ให้ประสิทธิภาพดีที่สุด

นอกจากนี้งานวิจัยซึ่งให้ผลตรงกันข้ามกับผลสรุป งานวิจัยของ ณัฐวุฒิ ศรีวิบูลย์ (2559) เรื่องการเปรียบเทียบประสิทธิภาพอัลกอริทึมเหมืองข้อมูลเพื่อวิเคราะห์ปัจจัยที่ส่งผลต่อการเกิดโรคมะเร็งโรคเบาหวาน (ชุดข้อมูลนี้ไม่ได้พิจารณาค่านอกเกณฑ์) พบว่าวิธีต้นไม้ตัดสินใจให้ค่าความถูกต้องสูงสุด R. Priya and P. Aruna (2012) เรื่องวิธีซัพพอร์ตเวกเตอร์แมชชีนและวิธีโครงข่ายประสาทเทียมในการวินิจฉัยโรคเบาหวานในจอประสาทตาโรคเบาหวาน (ชุดข้อมูลนี้ไม่ได้พิจารณาค่านอกเกณฑ์) พบว่าวิธีซัพพอร์ตเวกเตอร์แมชชีนให้ประสิทธิภาพดีที่สุด ทิพย์ธิดา วงศ์พิพันธ์ (2555) เรื่องการใช้เหมืองข้อมูลช่วยในการตัดสินใจการให้สินเชื่อโรคเบาหวาน (ชุดข้อมูลนี้ไม่ได้พิจารณาค่านอกเกณฑ์) พบว่าวิธีต้นไม้ตัดสินใจให้ค่าความถูกต้องมากที่สุด

5.3 ข้อเสนอแนะ

1) เพื่อให้ได้ข้อสรุปของผลการวิเคราะห์ข้อมูลที่มีความสมบูรณ์มากขึ้น ดังนั้นผู้วิจัยอาจวิเคราะห์ข้อมูลด้วยอัลกอริทึมประเภทอื่นๆ ได้แก่ วิธีฐานกฎ (Rule-Based) วิธี BayesNet วิธี Logistic วิธี SGD และวิธี VotedPerceptron

2) การวิเคราะห์ค่านอกเกณฑ์ ไม่มีหลักการณ์แบ่งที่แน่นอน ผู้วิจัยได้ทำการค้นหาข้อมูลจำนวนหนึ่งมาตรวจสอบค่านอกเกณฑ์ พบว่ามีค่าอยู่ระหว่าง 0-10 จึงแบ่งข้อมูลออกเป็น 3 ช่วง คือ น้อย ปานกลาง มาก ซึ่งอาจจะมีข้อมูลอื่นๆ ที่มีค่านอกเกณฑ์ไม่ได้อยู่ในช่วงดังกล่าว

3) ควรศึกษาวิธีการหาค่านอกเกณฑ์ อาจมีโปรแกรมอื่นๆ ที่มีประสิทธิภาพที่ดีกว่า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บรรณานุกรม

กิตติพล วิแสง, สิรภัทร เชี่ยวชาญวัฒนา และคำรณ สุนัติ, 2552, การวิเคราะห์ปัจจัยเสี่ยงของโรคเบาหวาน, การประชุมวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ (NCCIT) ครั้งที่ 5, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, กรุงเทพฯ, 8 น.

คณะแพทยศาสตร์ศิริราชพยาบาล สาขาศัลยศาสตร์ศีรษะ คอและเต้านม ภาควิชาศัลยศาสตร์, 2555, มะเร็งเต้านม เรื่องที่น่ารู้และข้อควรปฏิบัติสำหรับผู้ป่วยมะเร็งเต้านม, กรุงเทพฯ.

จิรา แก้วสุวรรณ์, 2549, การตรวจจับและการแก้ไขการวางตัวของภาพโดยใช้ซัพพอร์ตเวกเตอร์แมชชีน, วิทยานิพนธ์ปริญญาโท สาขาวิชาเทคโนโลยีคอมพิวเตอร์, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 55 น.

เดช ธรรมศิริ, วาทินี นุ้ยเพียร, ภัทราวุฒิ แสงศิริ, ภรัณยา อำมฤครัตน์, ณรงค์ โพธิ และพยุง มีสัจ, 2552, การให้คะแนนสินเชื่อโดยวิธีการทำเหมืองข้อมูลด้วยเทคนิคซับพอร์ตเวกเตอร์แมชชีน รวมทั้งการเลือกใช้ลักษณะที่เหมาะสมร่วมกับการหาค่าพารามิเตอร์ที่เหมาะสมด้วยวิธีค้นหาแบบกริช, การประชุมวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ (NCCIT) ครั้งที่ 5, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, กรุงเทพฯ, 11 น.

นิเวศ จิระวิชิตชัย, 2553, การค้นหาเทคนิคเหมืองข้อมูลเพื่อสร้างโมเดลการวิเคราะห์โรคอัตโนมัติ, มหาวิทยาลัยราชภัฏสวนสุนันทา, 131 น.

ทิพย์ธิดา วงศ์พิพันธ์, 2555, การใช้เหมืองข้อมูลช่วยในการตัดสินใจการให้สินเชื่อ, มหาบัณฑิต สาขาวิชาเทคโนโลยีคอมพิวเตอร์และการสื่อสาร คณะวิศวกรรมศาสตร์, มหาวิทยาลัยธุรกิจบัณฑิตย์, 86 น.

บทความอิเล็กทรอนิกส์, บัตรเครดิตคืออะไร, 2558, เข้าถึงได้ : http://money.sanook.com/323947/. สืบค้นวันที่ 23 ตุลาคม 2560.

บทความอิเล็กทรอนิกส์, Random Seed, 2560, เข้าถึงได้ : https://www.gravitechthai.com/detail_admin.php?masterkey=guru&cID=398. สืบค้นวันที่ 3 เมษายน 2561.

พยูน พาณิชย์กุล, 2548, การพัฒนาระบบดาต้าไมน์นิ่งโดยใช้ Decision Tree, โครงการพัฒนาระบบงานปริญญาวิทยาศาสตร์มหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ แขนงวิทยาการสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าคุณทหารลาดกระบัง.

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

พวงทอง ไกรพิบูลย์, 2557, เบาหวาน Diabetes mellitus, เข้าถึงได้ : http://haamor.com/th/. สืบค้นวันที่ 23 ตุลาคม 2560.

รุจิรา ธรรมสมบัติ, 2554, ระบบสนับสนุนการตัดสินใจในการเลือกใช้แพคเกจอินเทอร์เน็ตมือถือโดยใช้ต้นไม้ตัดสินใจ, สาขาคอมพิวเตอร์ธุรกิจ คณะบริหารธุรกิจ วิทยาลัยราชพฤกษ์, กรุงเทพฯ, 68 น.

วรพรรณ เจริญขำ, 2556, การตรวจสอบค่านอกเกณฑ์ในตัวอย่างสุ่มจากประชากรปกติ, วิทยาศาสตรมหาบัณฑิต (สถิติ), สถาบันบัณฑิตพัฒนบริหารศาสตร์, กรุงเทพฯ, 88 น.

วรรณสิริ ธุระชน, วรพจน์ สุเมธาวัฒนพงศ์ และณัฐวิภา ส่งสุข, 2557, ระบบการจำแนกพันธุ์ยางพาราโดยใช้ตัวจำแนกนาอีฟ เบย์, การประชุมวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ (NCCIT) ครั้งที่ 10, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, กรุงเทพฯ.

วาทินี นุ้ยเพียร, พยุง มีสัจ และเดช ธรรมศิริ, 2553, การเปรียบเทียบประสิทธิภาพและวิเคราะห์การจำแนกข้อมูลด้วยโครงข่ายประสาทเทียม ซัพพอร์ตเวกเตอร์แมชชีน นาอีฟ เบย์ และแครเนียรเรสต์เนเบอร์, น. 131-138, การประชุมวิชาการระดับชาติทางด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศครั้งที่ 5, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, กรุงเทพฯ.

สายชล สินสมบูรณ์ทอง, 2558, การทำเหมืองข้อมูล Data Mining, กรุงเทพฯ จามจุรี โปรดัก (จำกัด).

สายชล สินสมบูรณ์ทอง, 2560, การทำเหมืองข้อมูล Data Mining, ตีพิมพ์ครั้งที่ 2, กรุงเทพฯ จามจุรี โปรดัก (จำกัด).

Beckman, R.J., Cook, R.D., 1983, Outlier.....s. Journal of the Amarican Statistical Association and American Society for Quality. Vol. 25 ,No. 2 (May, 1983), pp.161-163. USA.

Berson, A., Stephen.J.Smith., 2001, Data Warehousing, Data Mining, & OLAP, Boston and McGraw-Hll, New York, USA.

Dr. WIlliam H. Wolberg., 1992. Breast Cancer Wisconsin (Original) Data Set, เข้าถึงได้ https://archive.ics.uci.edu/ml/datasets/breast+cancer+Wisconsin+(original). สืบค้นวันที่ 1 ตุลาคม 2560.

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Hawkins, 1980, Indentification of Outlier, Chapman and Holl, London.

I-Cheng Yeh. 2016. default of credit card clients Data Set. เข้าถึงได้ : https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients. สืบค้นวันที่ 1 ตุลาคม 2560.

National Institute of Diabetes and Digestive and Kidney Diseases. 2011. Pima Indians Diabetes Data Set. เข้าถึงได้ : https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes. สืบค้นวันที่ 1 ตุลาคม 2560.

Priya, R., Aruna, P., 2012, SVM and Neural Network based Diagnosis of Diabetic Retinopathy, International Journal of Computer Applications (0975 – 8887), Volume 41– No.1, March 2012.

Sriwiboon, N., 2016, A comparative efficiency of data mining algorithms for analysis of factors affecting the cancer, SNRU Journal of Science and Technology, 8(3), 344-352, Retrieved fromhttps://www.tcithaijo.org/index.php/snru_journal/article/view/72010.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R.B., 2001, Missing values estimation methods for DNA microarrays, Bioinformatics 17: 520-525.

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ภาคผนวก ก

รายละเอียดและตัวอย่างข้อมูลที่นำมาวิเคราะห์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

1. โรคมะเร็งเต้านมของรัฐวิสคอนซิน (Breast Cancer Wisconsin)

ตารางที่ ก-1 คุณลักษณะและรายละเอียดโรคมะเร็งเต้านมของรัฐวิสคอนซิน

คุณลักษณะ	รายละเอียด
X_1 ความหนาของก้อนเนื้อ (Clump Thickness)	หน่วย : มิลลิเมตร
X_2 ความสม่ำเสมอของขนาดเซลล์ (Uniformity of Cell Size)	หน่วย : นาโนเมตร
X_3 ความสม่ำเสมอของรูปร่างเซลล์ (Uniformity of Cell Shape)	หน่วย : ยูนิตเซลล์
X_4 การเกาะติดขอบของเซลล์ (Marginal Adhesion)	หน่วย : เซลล์
X_5 ขนาดเซลล์เดียว (Single Epithelial Cell Size)	หน่วย : มิลลิเมตร
X_6 นิวเคลียสไม่ถูกห่อหุ้ม (Bare Nuclei)	หน่วย : นิวเคลียส
X_7 ของเหลวในเซลล์ (Bland Chromatin)	หน่วย : นาโนเมตร
X_8 นิวคลีโอไลในภาวะปกติ (Normal Nucleoli)	หน่วย : เซลล์
X_9 การขยายตัวของเซลล์ (Mitoses)	หน่วย : เซลล์
Y ความเสี่ยงการเป็นโรคมะเร็งเต้านม (Class)	No คือไม่เป็นโรคมะเร็งเต้านม Yes คือเป็นโรคมะเร็งเต้านม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ก-2 ตัวอย่างข้อมูลโรคมะเร็งเต้านมของรัฐวิสคอนซิน

No	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	Y
1	5	1	1	1	2	1	3	1	1	no
2	5	4	4	5	7	10	3	2	1	no
3	3	1	1	1	2	2	3	1	1	no
4	6	8	8	1	3	4	3	7	1	no
5	4	1	1	3	2	1	3	1	1	no
6	8	10	10	8	7	10	9	7	1	yes
7	1	1	1	1	2	10	3	1	1	no
8	2	1	2	1	2	1	3	1	1	no
9	2	1	1	1	2	1	1	1	5	no
10	4	2	1	1	2	1	2	1	1	no
11	1	1	1	1	1	1	3	1	1	no
12	2	1	1	1	2	1	2	1	1	no
13	5	3	3	3	2	3	4	4	1	yes
14	1	1	1	1	2	3	3	1	1	no
15	8	7	5	10	7	9	5	5	4	yes
16	7	4	6	4	6	1	4	3	1	yes
17	4	1	1	1	2	1	2	1	1	no
18	4	1	1	1	2	1	3	1	1	no
19	10	7	7	6	4	10	4	1	2	yes
20	6	1	1	1	2	1	3	1	1	no
21	7	3	2	10	5	10	5	4	4	yes
22	10	5	5	3	6	7	7	10	1	yes
23	3	1	1	1	2	1	2	1	1	no
24	8	4	5	1	2	1	7	3	1	yes
25	1	1	1	1	2	1	3	1	1	no
26	5	2	3	4	2	7	3	6	1	yes
27	3	2	1	1	1	1	2	1	1	no
28	5	1	1	1	2	1	2	1	1	no
29	2	1	1	1	2	1	2	1	1	no

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2. โรคเบาหวานของชาวพิม่า ประเทศอินเดีย (Pima Indians Diabetes)

ตารางที่ ก-3 คุณลักษณะและรายละเอียดโรคเบาหวานของชาวพิม่า ประเทศอินเดีย

คุณลักษณะ	รายละเอียด
X_1 จำนวนครั้งของหญิงที่ตั้งครรภ์มาแล้ว (Number of times pregnant)	หน่วย : ครั้ง
X_2 ความเข้มข้นของกลูโคสในเลือด 2 ชั่วโมงก่อนการทดสอบระดับกลูโคสในช่องปาก (Plasma glucose concentration a 2 hours in an oral glucose tolerance test)	หน่วย : มิลลิกรัมต่อเดซิเบล
X_3 ความดันโลหิตสูง (Diastolic blood pressure)	หน่วย : มิลลิเมตรปรอท
X_4 ความหนาของชั้นไขมันใต้ผิวหนังบริเวณกล้ามที่ด้านหน้าของต้นแขน (Triceps skin fold thickness)	หน่วย : มิลลิเมตร
X_5 จำนวนอินซูลิน หลังจากอดอาหารภายใน 2 ชั่วโมง (2-Hour serum insulin)	หน่วย : 1,000,000 ยูนิตต่อมิลลิลิตร
X_6 ค่าดัชนีมวลกาย (Body mass index)	หน่วย : กิโลกรัมต่อตารางเมตร
X_7 พันธุกรรมของผู้ป่วยที่มีความเสี่ยงเป็นโรคเบาหวาน (Diabetes pedigree function)	หน่วย : ร้อยละ
X_8 อายุ (Age)	หน่วย : ปี
Y ความเสี่ยงการเป็นโรคเบาหวาน (Class)	No คือ ไม่เป็นโรคเบาหวาน Yes คือ เป็นโรคเบาหวาน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ก-4 ตัวอย่างข้อมูลโรคเบาหวานของชาวพิม่า ประเทศอินเดีย

No	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	Y
1	6	148	72	35	0	33.6	0.627	50	yes
2	1	85	66	29	0	26.6	0.351	31	no
3	8	183	64	0	0	23.3	0.672	32	yes
4	1	89	66	23	94	28.1	0.167	21	no
5	0	137	40	35	168	43.1	2.288	33	yes
6	5	116	74	0	0	25.6	0.201	30	no
7	3	78	50	32	88	31	0.248	26	yes
8	10	115	0	0	0	35.3	0.134	29	no
9	2	197	70	45	543	30.5	0.158	53	yes
10	8	125	96	0	0	0	0.232	54	yes
11	4	110	92	0	0	37.6	0.191	30	no
12	10	168	74	0	0	38	0.537	34	yes
13	10	139	80	0	0	27.1	1.441	57	no
14	1	189	60	23	846	30.1	0.398	59	yes
15	5	166	72	19	175	25.8	0.587	51	yes
16	7	100	0	0	0	30	0.484	32	yes
17	0	118	84	47	230	45.8	0.551	31	yes
18	7	107	74	0	0	29.6	0.254	31	yes
19	1	103	30	38	83	43.3	0.183	33	no
20	1	115	70	30	96	34.6	0.529	32	yes
21	3	126	88	41	235	39.3	0.704	27	no
22	8	99	84	0	0	35.4	0.388	50	no
23	7	196	90	0	0	39.8	0.451	41	yes
24	9	119	80	35	0	29	0.263	29	yes
25	11	143	94	33	146	36.6	0.254	51	yes
26	10	125	70	26	115	31.1	0.205	41	yes
27	7	147	76	0	0	39.4	0.257	43	yes
28	1	97	66	15	140	23.2	0.487	22	no
29	13	145	82	19	110	22.2	0.245	57	no

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3. การชำระเงินด้วยบัตรเครดิตของลูกค้า (Default of Credit Card Clients)

ตารางที่ ก-5 คุณลักษณะและรายละเอียดการชำระเงินด้วยบัตรเครดิตของลูกค้า

คุณลักษณะ	รายละเอียด
X_1 จำนวนเครดิตที่ได้รับ รวมทั้งสินเชื่อผู้บริโภคแต่ละรายและเครดิตครอบครัว (Amount of the given credit (NT dollar))	หน่วย : สกุลเงินไต้หวันแบบใหม่ มีหน่วยเป็น TWD
X_2 เพศ (Gender)	1 คือ เพศชาย 2 คือ เพศหญิง
X_3 การศึกษา (Education)	1 คือ สำเร็จการศึกษา 2 คือ กำลังศึกษาระดับมหาวิทยาลัย 3 คือ กำลังศึกษามัธยมศึกษาตอนปลาย 4 คือ อื่นๆ
X_4 สถานภาพการสมรส (Marital status)	1 คือ แต่งงาน 2 คือ โสด 3 คือ อื่นๆ
X_5 อายุ (Age)	หน่วย : ปี
X_6-X_{11} ประวัติการชำระเงินรายเดือนย้อนหลัง (ตั้งแต่เมษายนถึงกันยายน 2005) (History of past payment. We tracked the past monthly payment records (from April to September, 2005))	สถานการณ์ชำระเงินคืน เป็น -2 แทน ชำระเงินล่วงหน้า 1 เดือน -1 แทน ชำระเงินตรงเวลา 0 แทน ยังไม่ได้ชำระเงิน 1 แทน ชำระเงินล่าช้า 1 เดือน 2 แทน ชำระเงินล่าช้า 2 เดือน 3 แทน ชำระเงินล่าช้า 3 เดือน 4 แทน ชำระเงินล่าช้า 4 เดือน 5 แทน ชำระเงินล่าช้า 5 เดือน 6 แทน ชำระเงินล่าช้า 6 เดือน 7 แทน ชำระเงินล่าช้า 7 เดือนขึ้นไป
X_{12}-X_{17} ใบแจ้งจำนวนค่าใช้จ่าย (ตั้งแต่เมษายนถึงกันยายน 2005) (Amount of bill statement (NT dollar))	หน่วย : บาท

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ก-5 คุณลักษณะและรายละเอียดการชำระเงินด้วยบัตรเครดิตของลูกค้า (ต่อ)

คุณลักษณะ	รายละเอียด
X_{18}-X_{23} จำนวนเงินการชำระเงินก่อนหน้านี้ (ตั้งแต่เมษายนถึงกันยายน 2005) (Amount of previous payment (NT dollar))	หน่วย : บาท
Y การชำระเงินในเดือนถัดไป (Class)	No คือ ไม่ชำระ Yes คือ ชำระ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ก-6 ตัวอย่างข้อมูลการชำระเงินด้วยบัตรเครดิตของลูกค้า

No	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1	360000	2	1	1	57	-2	-2	-2	-2	-2
2	200000	2	1	1	59	0	0	0	0	0
3	200000	2	2	1	69	0	0	0	0	0
4	140000	2	3	2	61	1	-2	-2	-2	-2
5	360000	2	2	1	58	-2	-1	-1	-2	-2
6	180000	2	5	1	54	0	0	0	0	0
7	80000	2	2	1	55	0	0	0	0	0
8	390000	2	1	1	53	-1	-1	-2	-2	-1
9	440000	2	1	1	53	0	0	0	0	0
10	50000	2	2	1	55	-1	-1	-2	-2	-2
11	80000	2	2	2	60	-2	-2	-2	-2	-2
12	230000	2	2	1	54	-1	-1	2	2	-2
13	90000	2	3	1	53	0	0	0	0	0
14	80000	2	3	2	54	0	0	2	0	0
15	350000	2	3	1	64	0	0	0	0	0
16	440000	2	5	1	64	0	0	0	0	0
17	90000	2	3	1	54	3	2	2	2	2
18	40000	2	2	1	56	0	0	0	0	0
19	50000	2	2	1	61	3	2	0	0	0
20	50000	2	3	1	58	1	3	2	2	2
21	100000	2	3	1	55	-2	-2	-2	-2	-2
22	200000	2	3	1	56	0	0	0	0	0
23	180000	2	3	2	55	-1	0	0	0	0
24	180000	2	2	1	62	1	-2	-2	-2	-2
25	20000	2	2	1	58	-1	-1	-1	-1	-2
26	120000	2	3	1	63	0	0	0	0	0
27	80000	2	3	2	65	2	0	0	0	0
28	60000	2	3	1	66	0	0	2	2	4
29	140000	2	2	1	55	0	0	0	0	0

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ก-6 ตัวอย่างข้อมูลการชำระเงินด้วยบัตรเครดิตของลูกค้า (ต่อ)

No	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}	X_{19}	X_{20}
1	-2	-2	-2	-2	-2	1422	0	0	0	0
2	0	92215	94085	96174	98116	99490	101412	4300	4500	4438
3	0	45870	46907	47952	49052	49936	50977	2100	2100	2200
4	-2	0	0	0	0	0	0	0	0	0
5	-2	0	600	0	0	0	1459	600	0	0
6	0	87695	88545	90403	92245	94082	96050	3200	3279	3343
7	0	79816	29536	28870	25049	23881	23668	1600	1514	1505
8	0	611	0	0	0	343	1343	0	0	0
9	0	295180	284145	255143	234395	216318	202530	13007	9102	10003
10	-2	1100	0	0	0	0	0	0	0	0
11	-2	577	494	438	387	246	939	500	438	387
12	-2	187	3327	3100	0	0	0	3327	0	0
13	0	60769	61943	62082	61436	27507	29218	2207	2500	1479
14	0	16613	19828	19227	20593	20951	21460	3500	0	2000
15	0	30726	27639	28890	31090	32265	33737	3000	2000	3000
16	0	133996	136606	136771	141131	147251	150332	6319	5000	6625
17	0	86228	85437	90635	91745	87466	85662	1500	7500	3500
18	0	30048	31049	32144	33157	33657	34359	1805	1900	1851
19	0	52659	51213	51696	50422	30145	30203	4	2346	2155
20	2	39649	40194	41150	41670	42188	43360	1500	1900	1500
21	-2	390	390	390	390	390	390	390	390	390
22	0	193392	159546	158486	160375	163444	0	5039	8000	3500
23	0	52824	53165	54842	54868	56286	56621	2000	3002	3000
24	-2	0	0	0	0	0	0	0	0	0
25	-1	17895	13576	18286	-200	0	18486	13605	19352	0
26	0	49394	51261	51333	51174	44680	43760	3000	2000	2000
27	0	85882	84985	84247	81619	30095	20639	3900	4000	2048
28	4	47350	51230	54784	58232	61670	65120	5000	5000	5000
29	0	108038	110216	84516	72391	61298	62193	4200	2822	2336

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ก-6 ตัวอย่างข้อมูลการชำระเงินด้วยบัตรเครดิตของลูกค้า (ต่อ)

No	X_{21}	X_{22}	X_{23}	Y
1	1424	0	6624	no
2	3900	3800	3500	no
3	2000	2000	2011	no
4	0	0	0	yes
5	0	1459	0	no
6	3366	3483	3507	no
7	1005	2026	2000	no
8	343	1000	1076	no
9	5000	5000	6000	no
10	0	0	0	no
11	246	939	350	no
12	0	0	0	no
13	1000	3473	2500	no
14	1000	1000	1000	no
15	2000	2000	2000	no
16	10000	5443	5501	no
17	0	3300	3500	yes
18	1351	1394	1464	yes
19	1046	1073	1246	yes
20	1500	2000	1500	no
21	390	390	390	yes
22	3269	0	0	no
23	2300	2345	2000	no
24	0	0	0	yes
25	200	18486	5656	no
26	2000	2000	2000	no
27	1300	1700	401	no
28	5000	5000	1000	no
29	2588	2250	2491	no

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตัวอย่างการแบ่งข้อมูล

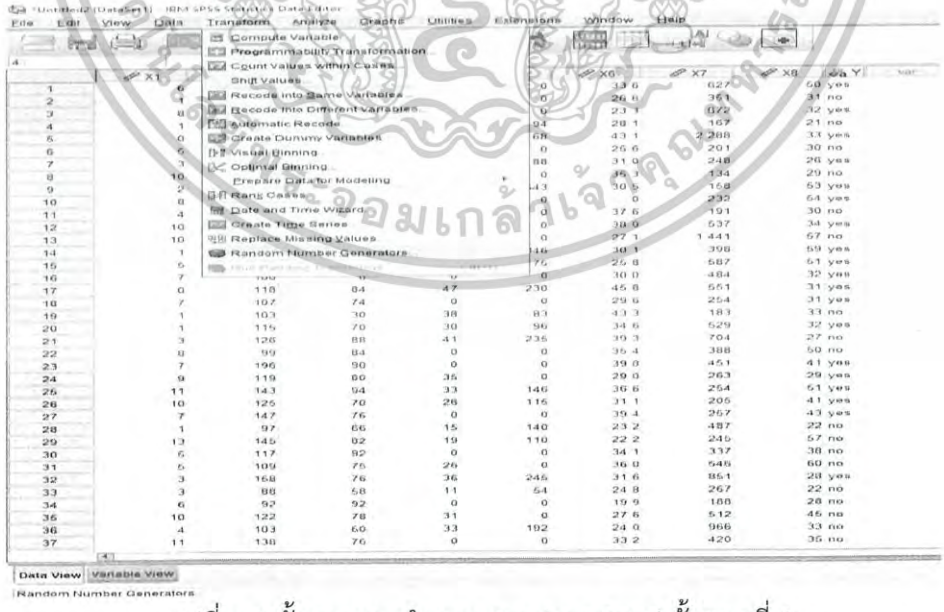
โรคเบาหวานของชาวพิม่า ประเทศอินเดีย (Pima Indians Diabetes) จำนวนข้อมูลทั้งหมด 768 ค่า ทำการแบ่งข้อมูลด้วยโปรแกรม SPSS กำหนด Random Seed 10

1.) แบ่งข้อมูลออกเป็น 70% และ 30% จากข้อมูลทั้งหมด

	X1	X2	X3	X4	X5	X6	X7	X8	Y
1	6	148	72	35	0	33.6	627	50	yes
2	1	85	66	29	0	26.6	351	31	no
3	8	183	64	0	0	23.3	672	32	yes
4	1	89	66	23	94	28.1	167	21	no
5	0	137	40	35	168	43.1	2.288	33	yes
6	5	116	74	0	0	25.6	201	30	no
7	3	78	50	32	88	31.0	248	26	yes
8	10	115	0	0	0	35.3	134	29	no
9	2	197	70	45	543	30.5	158	53	yes
10	8	125	96	0	0	0	232	54	yes
11	4	110	92	0	0	37.6	191	30	no
12	10	168	74	0	0	38.0	537	34	yes
13	10	139	80	0	0	27.1	1.441	57	no
14	1	189	60	23	846	30.1	398	59	yes
15	5	166	72	19	175	25.8	587	51	yes
16	7	100	0	0	0	30.0	484	32	yes
17	0	118	84	47	230	45.8	551	31	yes
18	7	107	74	0	0	29.6	254	31	yes
19	1	103	30	38	83	43.3	183	33	no
20	1	115	70	30	96	34.6	529	32	yes
21	3	126	88	41	235	39.3	704	27	no
22	8	99	84	0	0	35.4	388	50	no
23	7	196	90	0	0	39.8	451	41	yes
24	9	119	80	35	0	29.0	263	29	yes
25	11	143	94	33	146	36.6	254	51	yes
26	10	125	70	26	115	31.1	205	41	yes
27	7	147	76	0	0	39.4	257	43	yes
28	1	97	66	15	140	23.2	487	22	no
29	13	145	82	19	110	22.2	245	57	no
30	5	117	92	0	0	34.1	337	38	no
31	5	109	75	26	0	36.0	546	60	no
32	3	158	76	36	245	31.6	851	28	yes
33	3	88	58	11	54	24.8	267	22	no
34	6	92	92	0	0	19.9	188	28	no
35	10	122	78	31	0	27.6	512	45	no
36	4	103	60	33	192	24.0	966	33	no
37	11	138	76	0	0	33.2	420	35	no

รูปที่ ก-1 ข้อมูลโรคเบาหวานของชาวพิม่า ประเทศอินเดีย

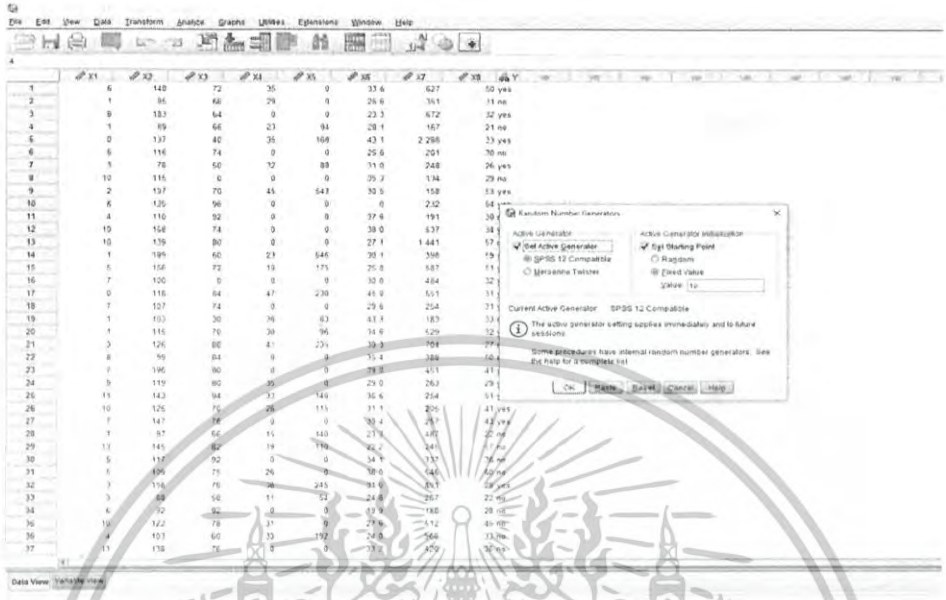
1.1 กำหนดค่า Random Seed → คลิกที่ Transform → Random Number Generators



รูปที่ ก-2 ขั้นตอนการกำหนด Random Seed ขั้นตอนที่ 1

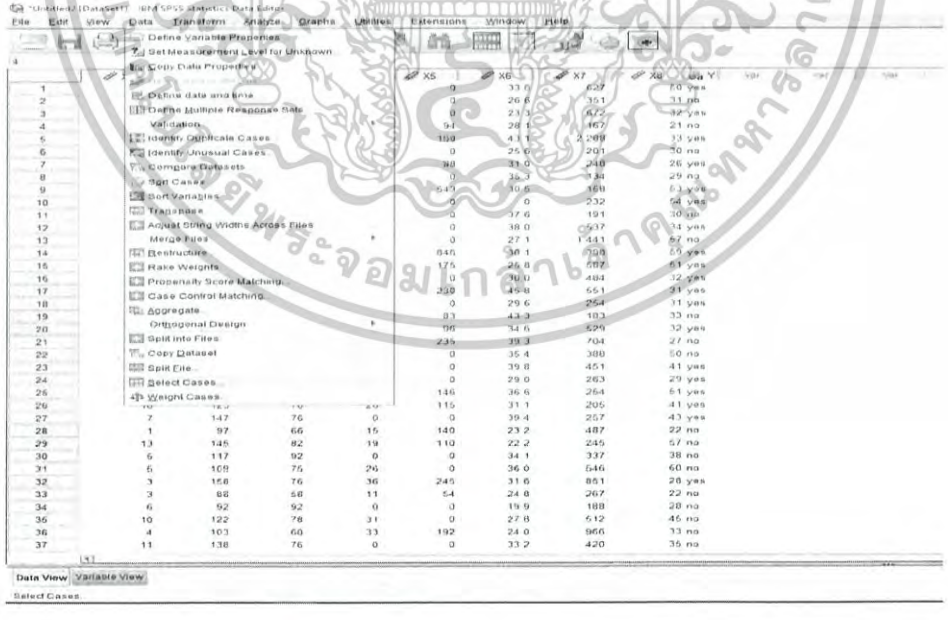
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

1.2 คลิกที่ Set Active Generator → คลิกที่ Set Starting Point → ใส่ค่า → Value 10 → ok



รูปที่ ก-3 ขั้นตอนการกำหนด Random Seed ขั้นตอนที่ 2

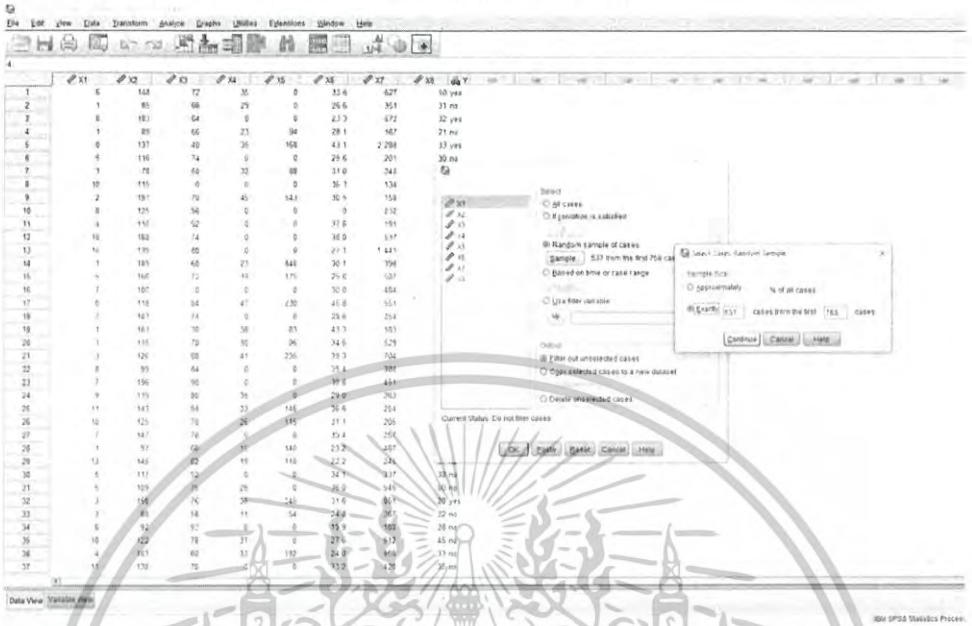
1.3 ทำการสุ่มข้อมูล 70:30 หลังจากการกำหนด Random Seed → คลิก Data → Select Cases



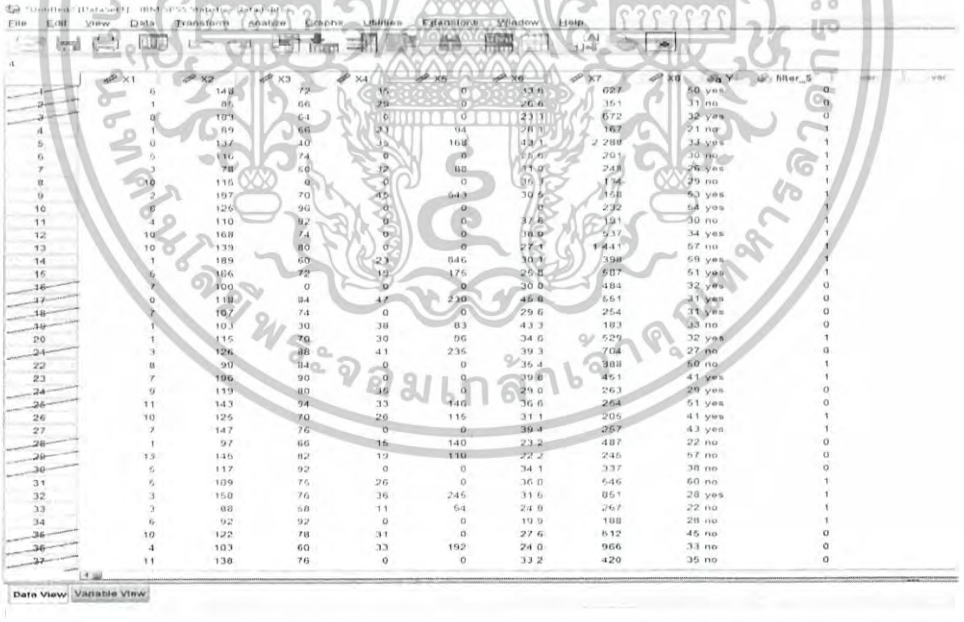
รูปที่ ก-4 ขั้นตอนการกำหนด Random Seed ขั้นตอนที่ 3

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

1.4 คลิก Random sample of case ⟶ คลิก Sample ⟶ ใส่ค่า Exactly 537 cases from the first 768 cases ⟶ Continue ⟶ OK



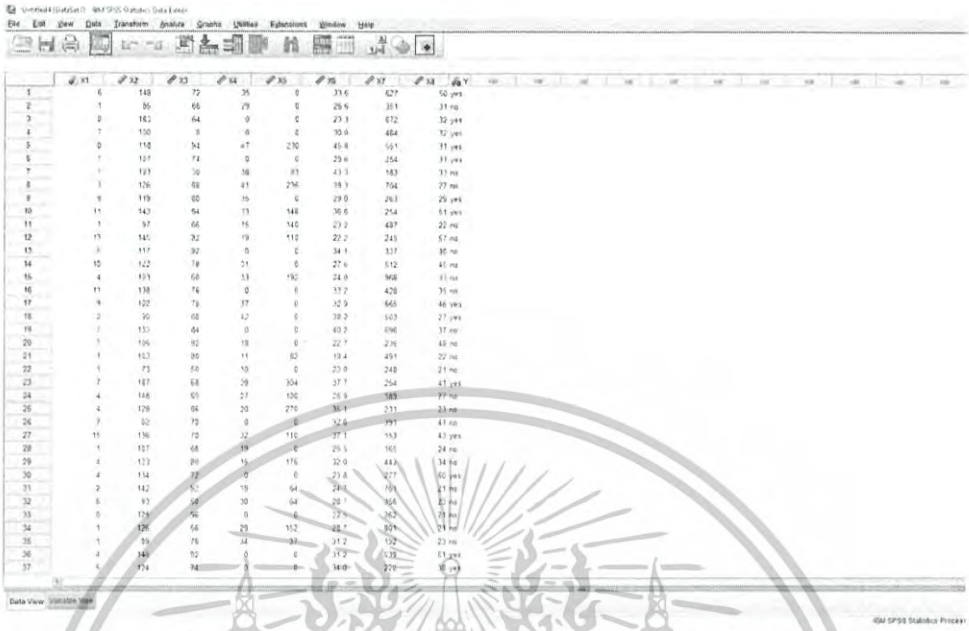
รูปที่ ก-5 ขั้นตอนการกำหนด Random Seed ขั้นตอนที่ 4



รูปที่ ก-6 ผลการแบ่งข้อมูล 70:30 โรคเบาหวานของชาวพิม่า ประเทศอินเดีย ด้วยโปรแกรม SPSS

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.) นำข้อมูล 30% มาแบ่งต่อเป็น 20% และ 10%



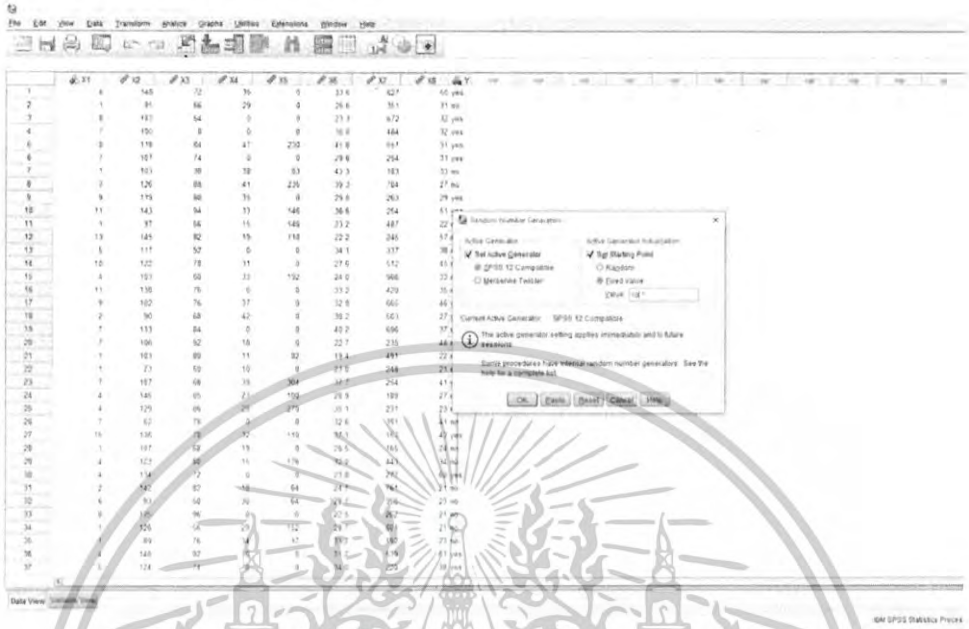
รูปที่ ก-7 ข้อมูลโรคเบาหวานของชาวพิม่า ประเทศอินเดียจำนวน 30%

2.1 กำหนด Random Seed ▶ คลิกที่ Transform ▶ Random Number Generators

รูปที่ ก-8 ขั้นตอนการแบ่งข้อมูล ขั้นตอนที่ 1

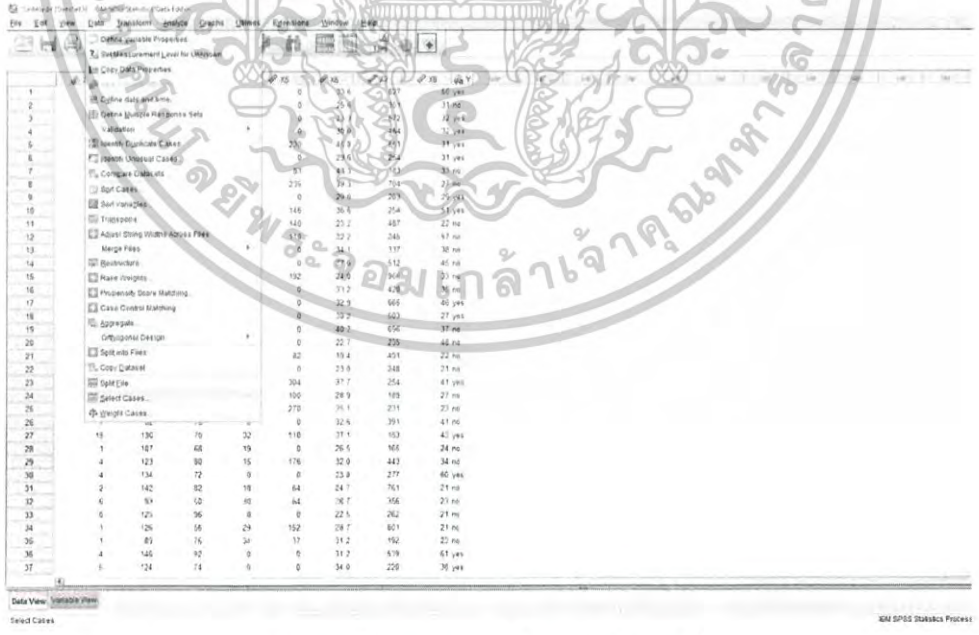
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.2 คลิกที่ Set Active Generator → คลิกที่ Set Starting Point → ใส่ค่า Value 10 → ok



รูปที่ ก-9 ขั้นตอนการแบ่งข้อมูล ขั้นตอนที่ 2

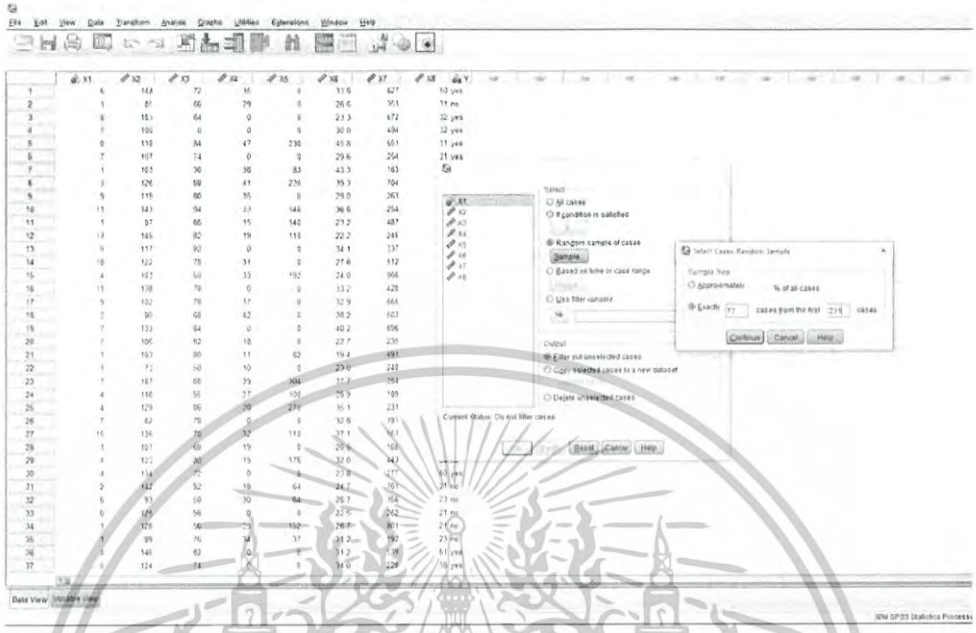
2.3 ทำการสุ่มข้อมูล 20:10 หลังจากการกำหนด Random Seed → คลิก Data → Select Cases



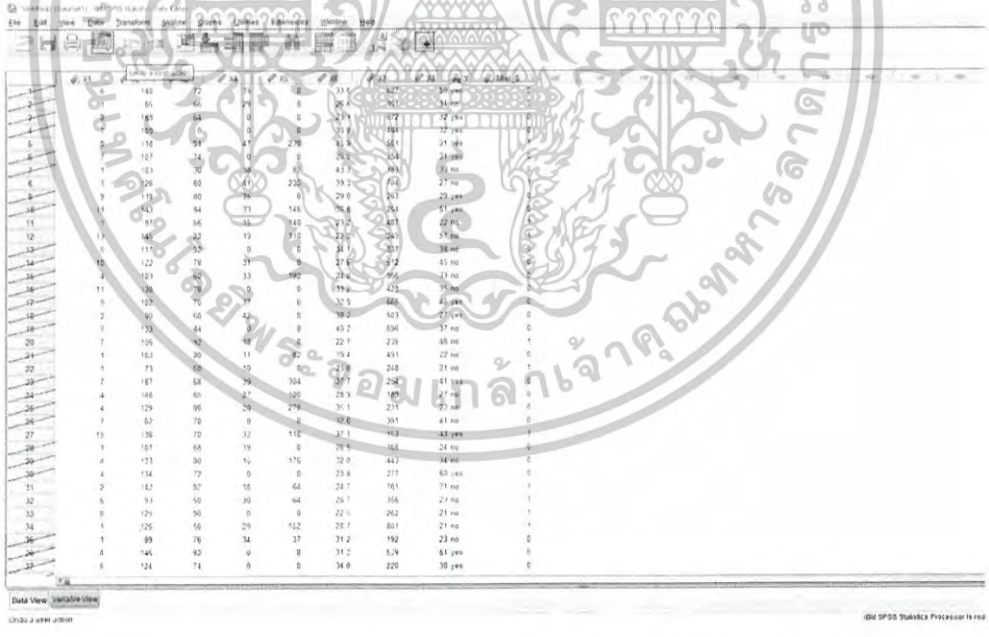
รูปที่ ก-10 ขั้นตอนการแบ่งข้อมูล ขั้นตอนที่ 3

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.4 คลิก Random sample of case ⟶ คลิก Sample ⟶ ใส่ค่า Exactly 77 cases from the first 231 cases ⟶ Continue ⟶ OK



รูปที่ ก-11 ขั้นตอนการแบ่งข้อมูล ขั้นตอนที่ 4



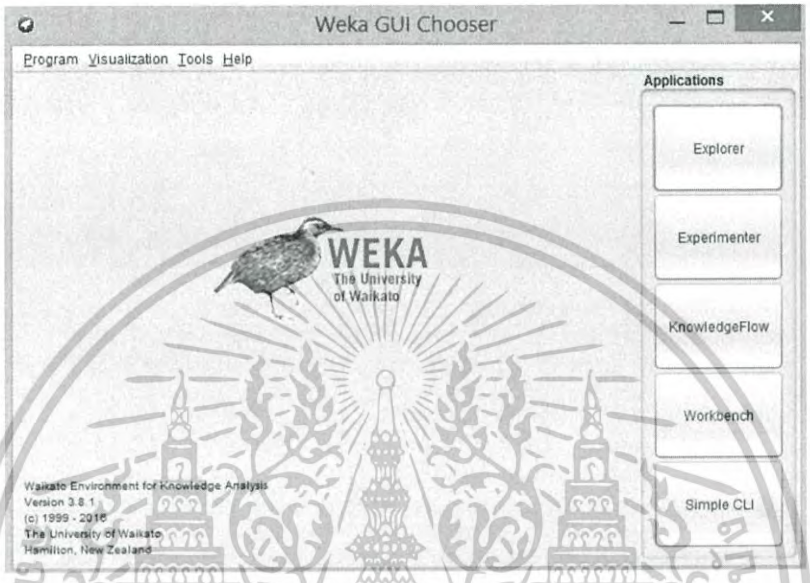
รูปที่ ก-12 ผลการแบ่งข้อมูล 20:10 โรคเบาหวานของชาวพิม่า ประเทศอินเดีย ด้วยโปรแกรม SPSS

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

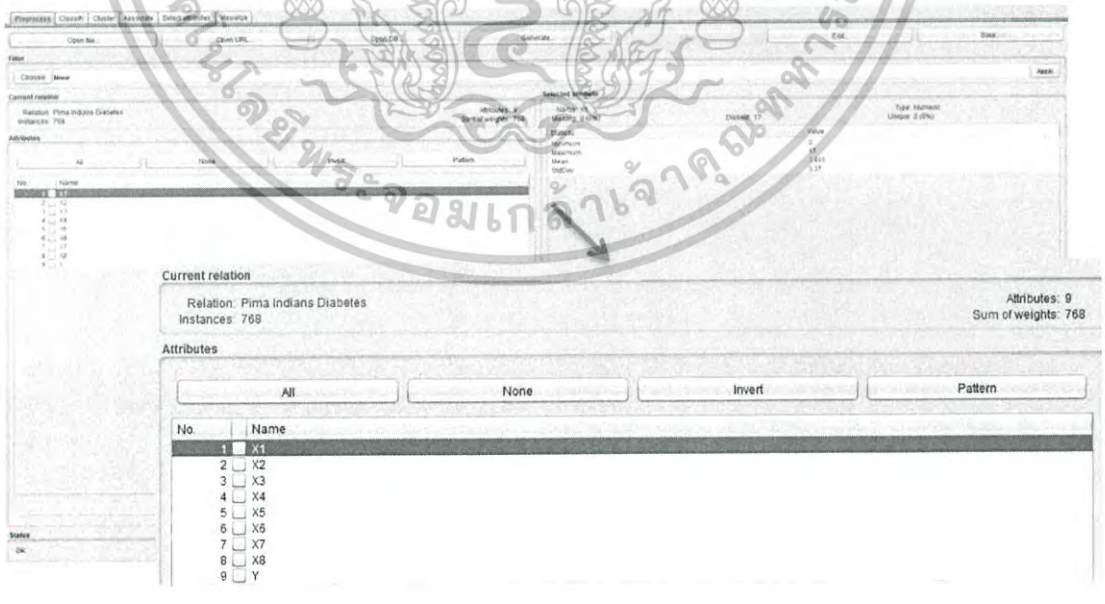
ตัวอย่างการแบ่งข้อมูล

โรคเบาหวานของชาวพิม่า ประเทศอินเดีย (Pima Indians Diabetes) จำนวนข้อมูลทั้งหมด 768 ค่า ทำการแบ่งข้อมูลด้วยโปรแกรม WEKA กำหนด Random Seed 10

1.) เปิด WEKA Explorer Open File (นามสกุล .csv) ไฟล์ข้อมูล จะได้ข้อมูล โรคเบาหวานของชาวพิม่า ประเทศอินเดีย (Pima Indians Diabetes) จำนวนข้อมูลทั้งหมด 768 ค่า



รูปที่ ก-13 หน้าต่างโปรแกรม WEKA



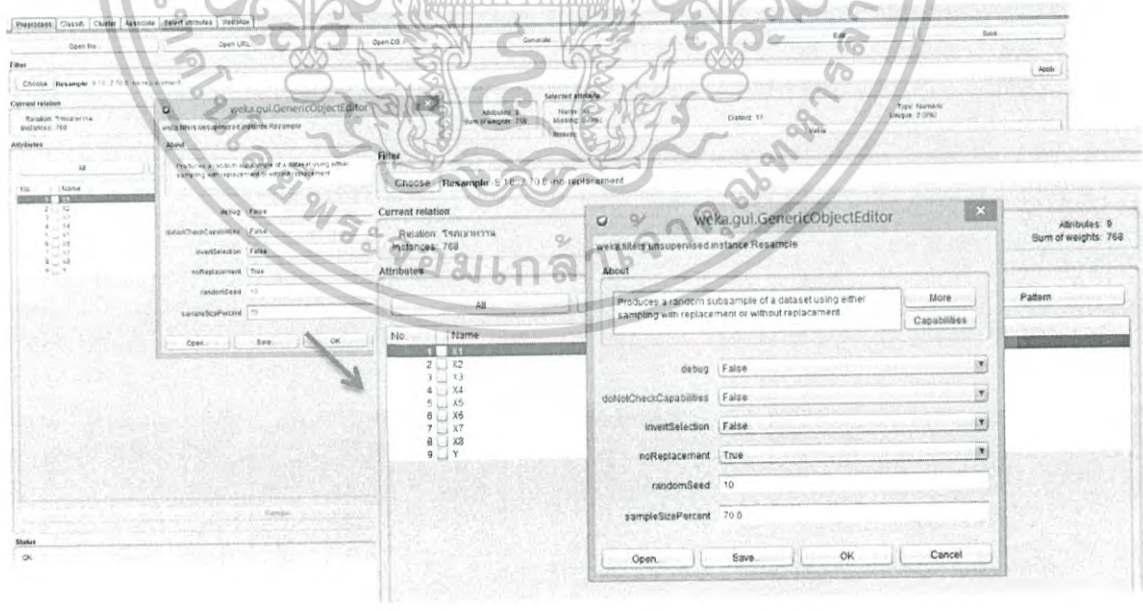
รูปที่ ก-14 แสดงจำนวนข้อมูล โรคเบาหวานของชาวพิม่า ประเทศอินเดีย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.) กดปุ่ม Choose → filters → unsupervised → instance → Resample

รูปที่ ก-15 ขั้นตอนการแบ่งข้อมูล การเลือกการสุ่ม

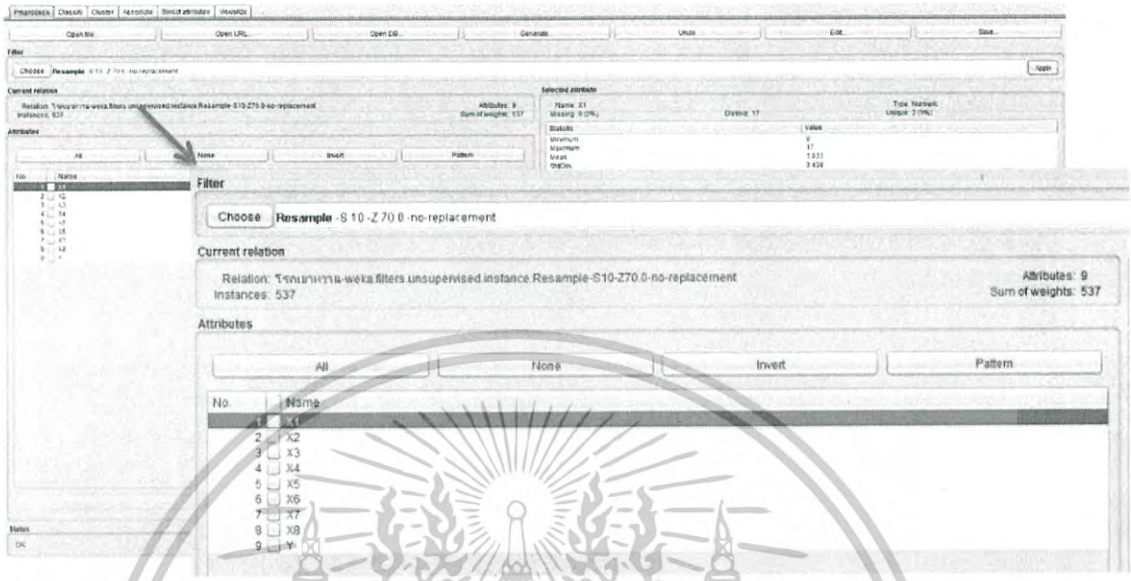
3.) กดช่อง Resample - S 1 -Z 100.0 → ในช่อง invertSelection เป็น False → ในช่อง noReplacement เป็น True → กำหนดค่า randomSeed เท่ากับ 10 → ในช่อง sampleSizePercent กำหนดให้เป็น 70 คือ 70%



รูปที่ ก-16 ขั้นตอนการแบ่งข้อมูล 70%

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

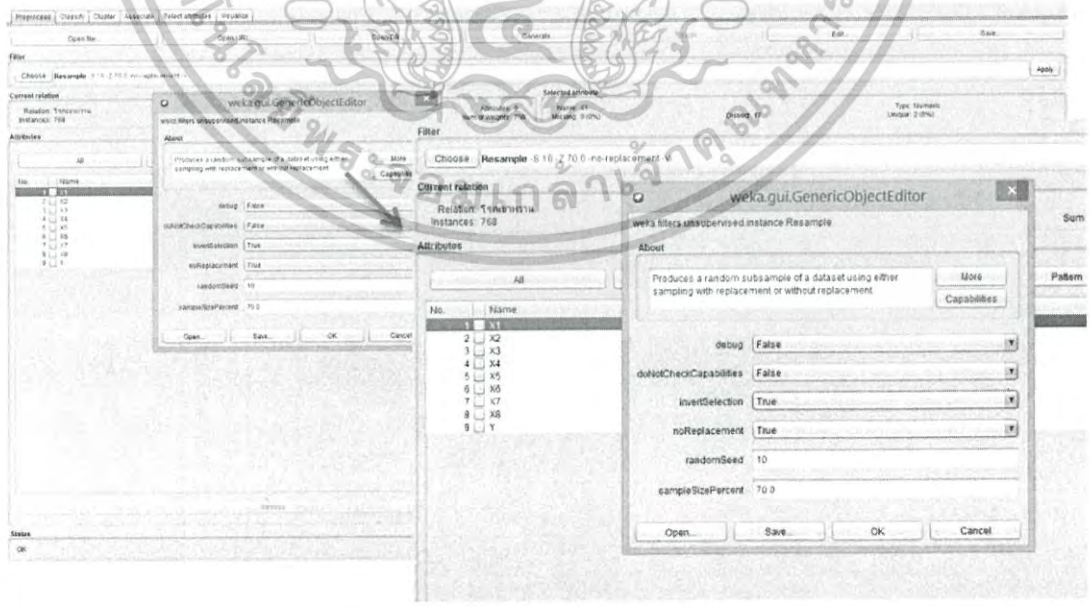
4.) กด Apply จะแสดงค่าข้อมูล 70% จำนวน 537 ค่า ⟶ กด Save กำหนดชื่อไฟล์เป็น Pima 70 กด ⟶ Undo



รูปที่ ก-17 แสดงผลลัพธ์ข้อมูล 70%

5.) กดช่อง Resample - S 10 ⟶ ในช่อง invertSelection เป็น True ⟶ ในช่อง noReplacement เป็น True ⟶ กำหนดค่า randomSeed เท่ากับ 10 ⟶ ในช่อง sampleSizePercent กำหนดให้เป็น 70 เพื่อต้องการแบ่งข้อมูล 30% จาก 70%

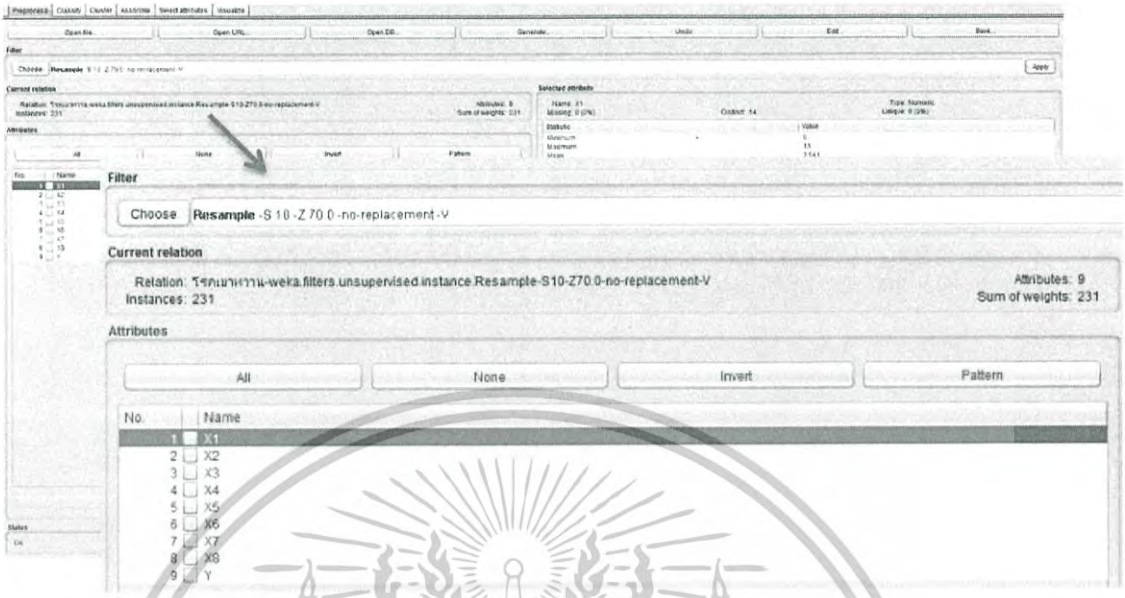
6.)



รูปที่ ก-18 ขั้นตอนการแบ่งข้อมูล 30

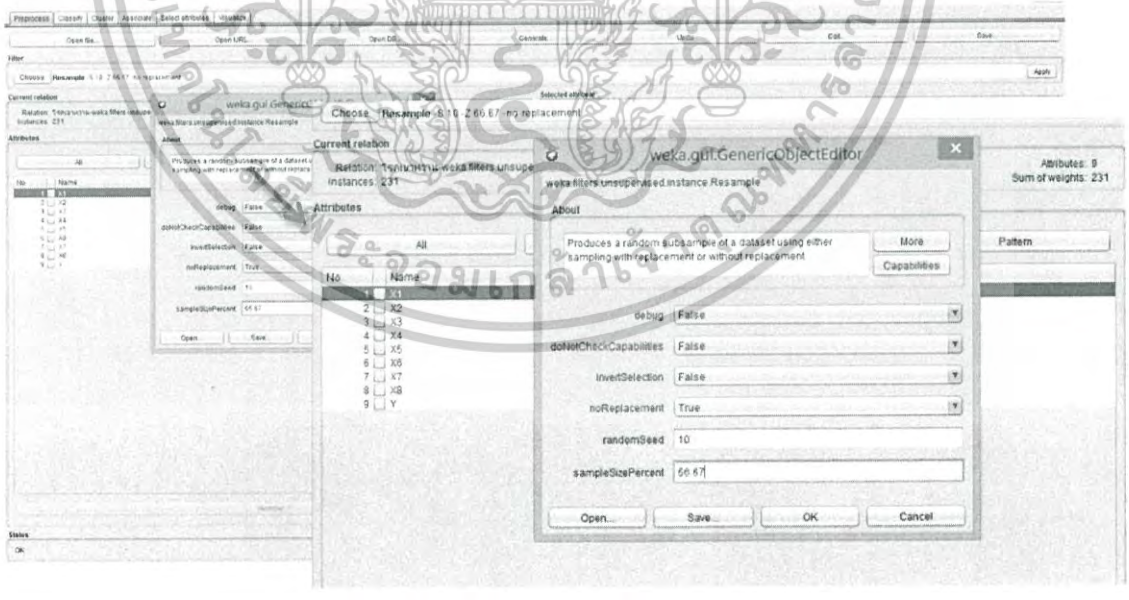
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

7.) กด Apply จะได้ข้อมูล 30% จำนวน 231 ค่า



รูปที่ ก-19 แสดงผลลัพธ์ข้อมูล 30%

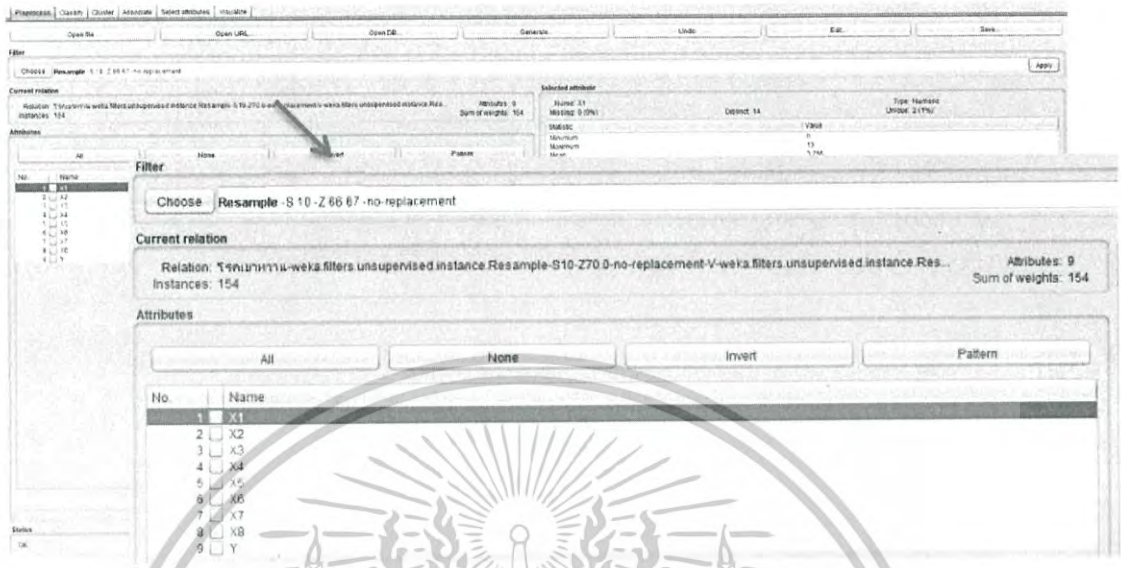
8.) กดช่อง Resample - S 10 → ในช่อง invertSelection เป็น False → ในช่อง noReplacement เป็น True → กำหนดค่า randomSeed เท่ากับ 10 → ในช่อง sampleSizePercent กำหนดให้เป็น 66.67 เพื่อต้องการแบ่งข้อมูล 20% จาก 30%



รูปที่ ก-20 ขั้นตอนการแบ่งข้อมูล 20%

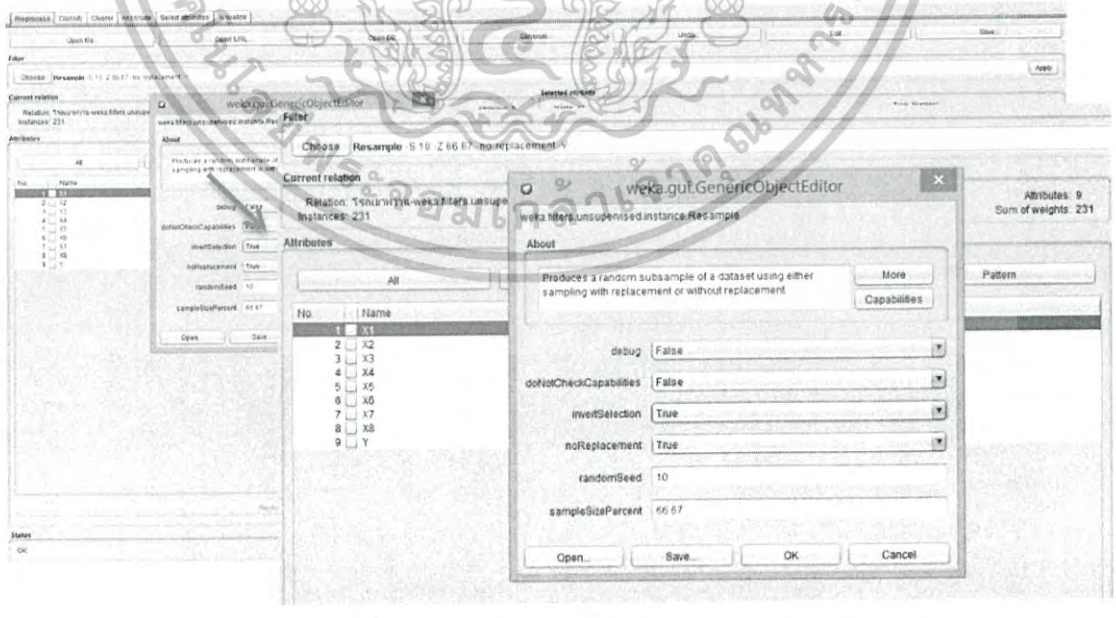
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

9.) กด Apply จะได้ข้อมูล 20% จำนวน 154 ค่า ⟶ กด Save กำหนดชื่อไฟล์เป็น Pima 20 ⟶ กด Undo



รูปที่ ก-21 แสดงผลลัพธ์ข้อมูล 20%

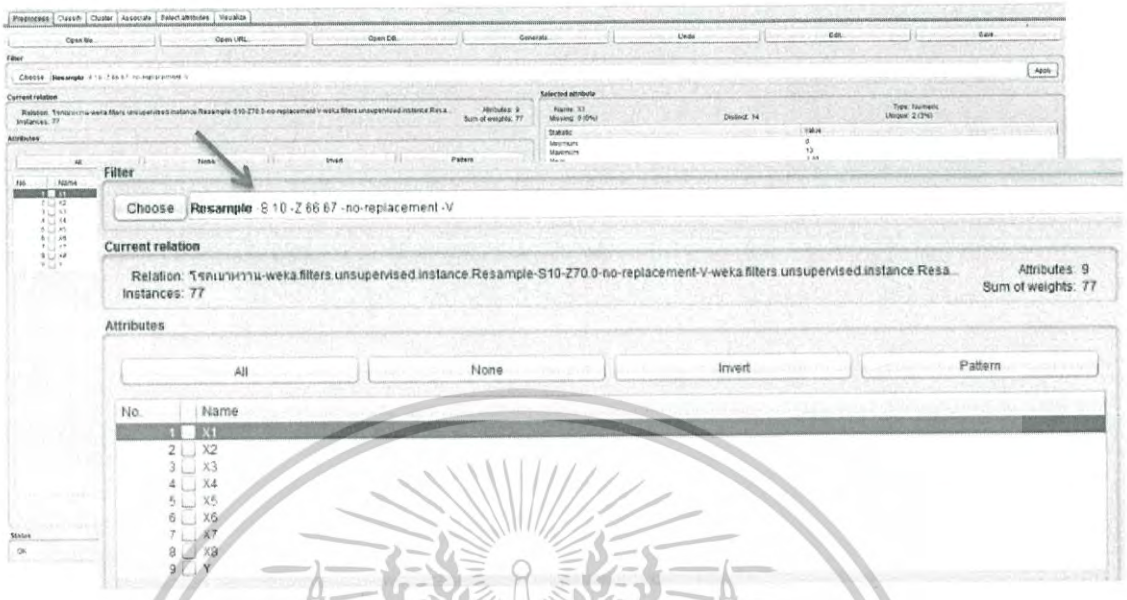
10.)กดช่อง Resample - S 10 ⟶ ในช่อง invertSelection เป็น True ⟶ ในช่อง noReplacement เป็น True ⟶ กำหนดค่า randomSeed เท่ากับ 10 ⟶ ในช่อง sampleSizePercent กำหนดให้เป็น 66.67 เพื่อต้องการแบ่งข้อมูล 10% จาก 30%



รูปที่ ก-22 ขั้นตอนการแบ่งข้อมูล 10%

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

11.) กด Apply จะได้ข้อมูล 10% จำนวน 77 ค่า ⟶ กด Save กาหนดชื่อไฟล์เป็น Pima 10



รูปที่ ก-23 แสดงผลลัพธ์ข้อมูล 10%

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

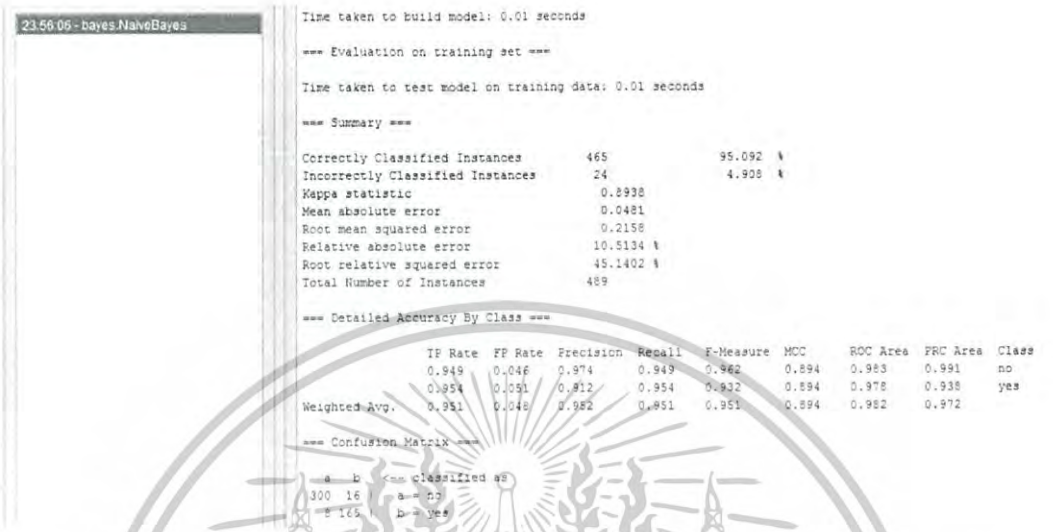
ภาคผนวก ข

การวิเคราะห์ข้อมูล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ข้อมูลโรคมะเร็งเต้านมของรัฐวิสคอนซิน (Breast Cancer Wisconsin)

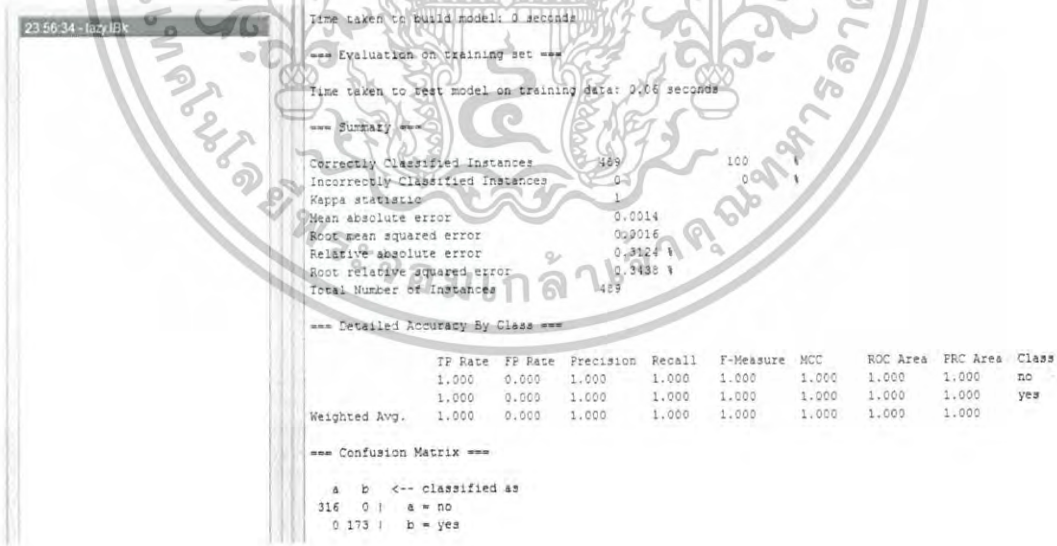
1. วิธีนาอีฟเบส์



```
23:56:06 - bayes.NaiveBayes

Time taken to build model: 0.01 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances         465               95.092 %
Incorrectly Classified Instances        24                4.908 %
Kappa statistic                          0.8938
Mean absolute error                      0.0481
Root mean squared error                  0.2158
Relative absolute error                 10.5134 %
Root relative squared error             45.1402 %
Total Number of Instances              489

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.949    0.046    0.974      0.949   0.962      0.894    0.983     0.991     no
                 0.954    0.051    0.912      0.954   0.932      0.894    0.978     0.938     yes
Weighted Avg.    0.951    0.048    0.952      0.951   0.951      0.894    0.982     0.972

=== Confusion Matrix ===

   a   b   <-- classified as
 300  16 |   a = no
   8 165 |   b = yes
```

รูปที่ ข-1 ผลการวิเคราะห์ข้อมูลสำหรับการสร้างตัวแบบวิธีนาอีฟเบส์

2. วิธีเพื่อนบ้านใกล้สุด k ตัว

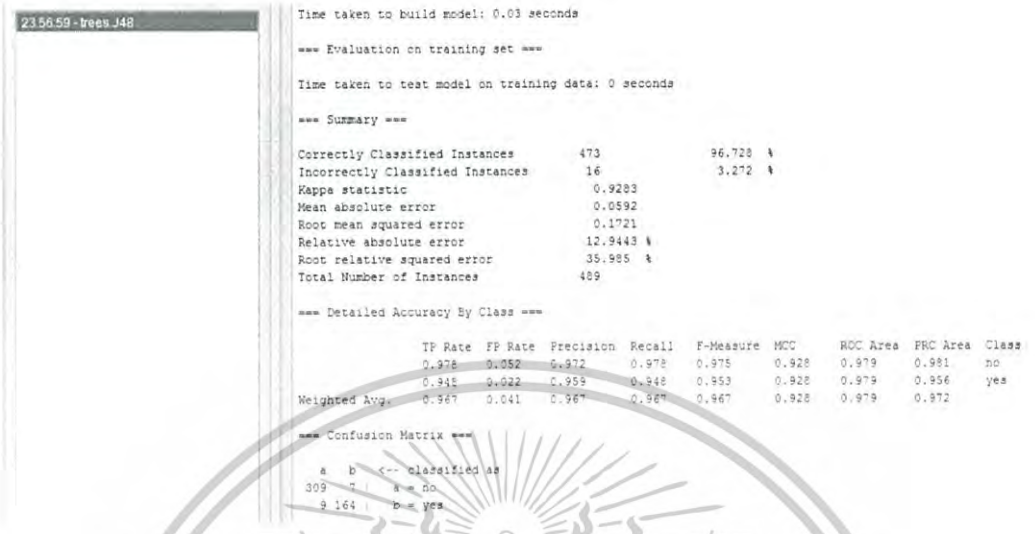


```
23:56:34 - lazy.IBk

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.06 seconds

=== Summary ===

Correctly Classified Instances         489               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0014
Root mean squared error                  0.0016
Relative absolute error                  0.3124 %
Root relative squared error              0.3438 %
Total Number of Instances              489

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000   1.000      1.000    1.000     1.000     no
                 1.000    0.000    1.000      1.000   1.000      1.000    1.000     1.000     yes
Weighted Avg.    1.000    0.000    1.000      1.000   1.000      1.000    1.000     1.000

=== Confusion Matrix ===

   a   b   <-- classified as
 316   0 |   a = no
   0 173 |   b = yes
```

รูปที่ ข-2 ผลการวิเคราะห์ข้อมูลสำหรับการสร้างตัวแบบวิธีเพื่อนบ้านใกล้สุด k ตัว

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3. วิธีต้นไม้ตัดสินใจ



```
23:56:59 - trees.J48

Time taken to build model: 0.03 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances         473               96.728 %
Incorrectly Classified Instances        16                3.272 %
Kappa statistic                          0.9283
Mean absolute error                      0.0592
Root mean squared error                  0.1721
Relative absolute error                 12.9443 %
Root relative squared error             35.985 %
Total Number of Instances              489

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.978     0.052     0.972     0.978     0.975     0.928     0.979     0.981     no
               0.948     0.022     0.959     0.948     0.953     0.928     0.979     0.956     yes
Weighted Avg.  0.967     0.041     0.967     0.967     0.967     0.928     0.979     0.972

=== Confusion Matrix ===

   a   b   <-- classified as
 309   7 |   a = no
   9 164 |   b = yes
```

รูปที่ ข-3 ผลการวิเคราะห์ข้อมูลสำหรับการสร้างตัวแบบวิธีต้นไม้ตัดสินใจ

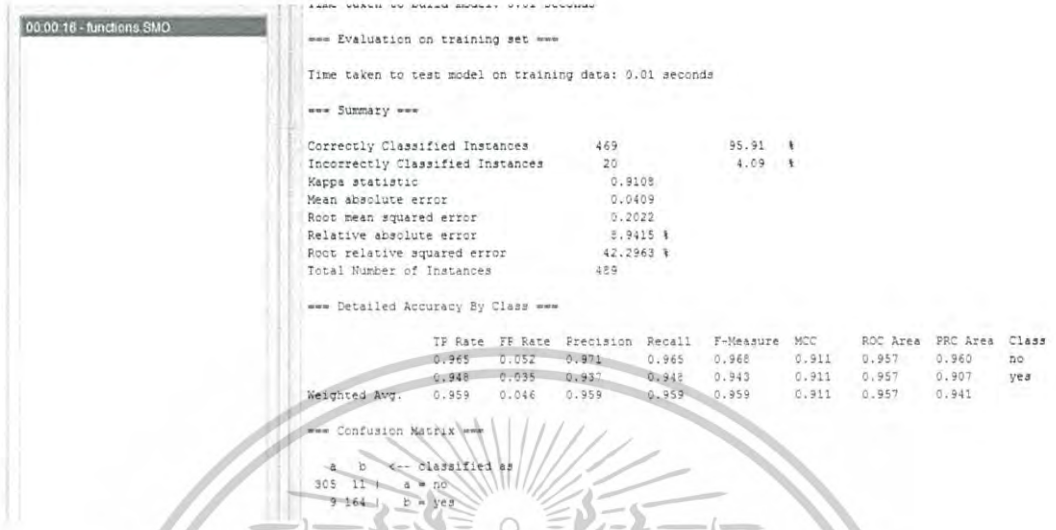
4. วิธีโครงข่ายประสาทเทียม

```
23:58:02 - functions.MultilayerPerceptron

Time taken to build model: 6.62 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances         476               97.3415 %
Incorrectly Classified Instances        13                2.6585 %
Kappa statistic                          0.9424
Mean absolute error                      0.0448
Root mean squared error                  0.1596
Relative absolute error                  9.7955 %
Root relative squared error             33.3721 %
Total Number of Instances              489

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.968     0.017     0.990     0.968     0.979     0.943     0.983     0.988     no
               0.983     0.032     0.944     0.983     0.963     0.943     0.984     0.963     yes
Weighted Avg.  0.973     0.022     0.974     0.973     0.974     0.943     0.984     0.979

=== Confusion Matrix ===

   a   b   <-- classified as
 306  10 |   a = no
   3 170 |   b = yes
```

รูปที่ ข-4 ผลการวิเคราะห์ข้อมูลสำหรับการสร้างตัวแบบวิธีโครงข่ายประสาทเทียม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

5. วิธีซัพพอร์ทเวกเตอร์แมชชีน

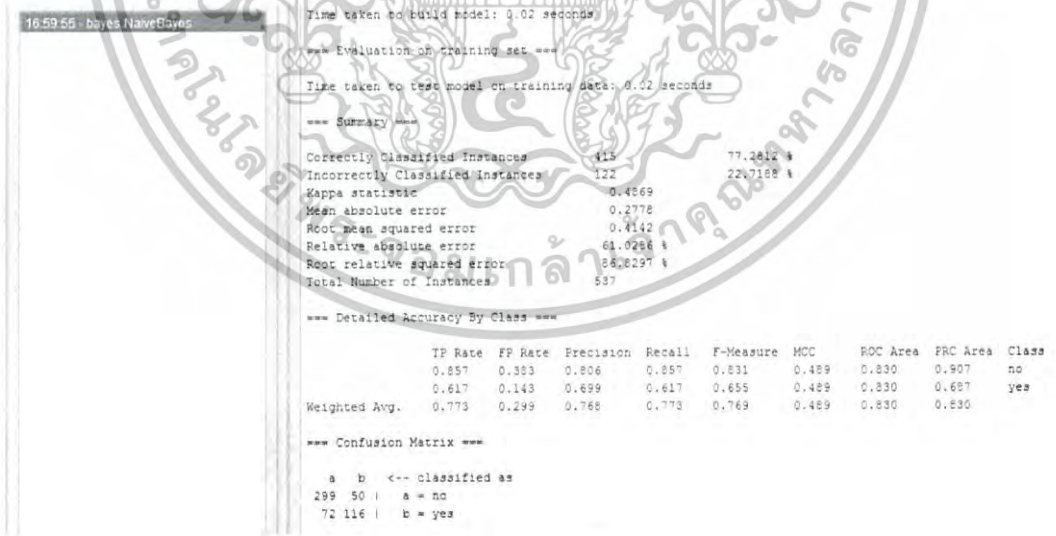


```
00:00:16 - functions.SMO

=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances         469               95.91   %
Incorrectly Classified Instances        20                4.09   %
Kappa statistic                          0.9108
Mean absolute error                      0.0409
Root mean squared error                  0.2022
Relative absolute error                  8.9415 %
Root relative squared error             42.2963 %
Total Number of Instances              489

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.965    0.052    0.971      0.965    0.968      0.911    0.957     0.960     no
                 0.948    0.035    0.937      0.948    0.943      0.911    0.957     0.907     yes
Weighted Avg.    0.959    0.046    0.959      0.959    0.959      0.911    0.957     0.941

=== Confusion Matrix ===

   a   b   <-- classified as
 305  11 |   a = no
   9 164 |   b = yes
```

รูปที่ ข-5 ผลการวิเคราะห์ข้อมูลสำหรับการสร้างตัวแบบวิธีซัพพอร์ทเวกเตอร์แมชชีน

ข้อมูลโรคเบาหวานของชาวพิม่า ประเทศอินเดีย (Pima Indians Diabetes)

1. วิธีนาอีฟเบส์

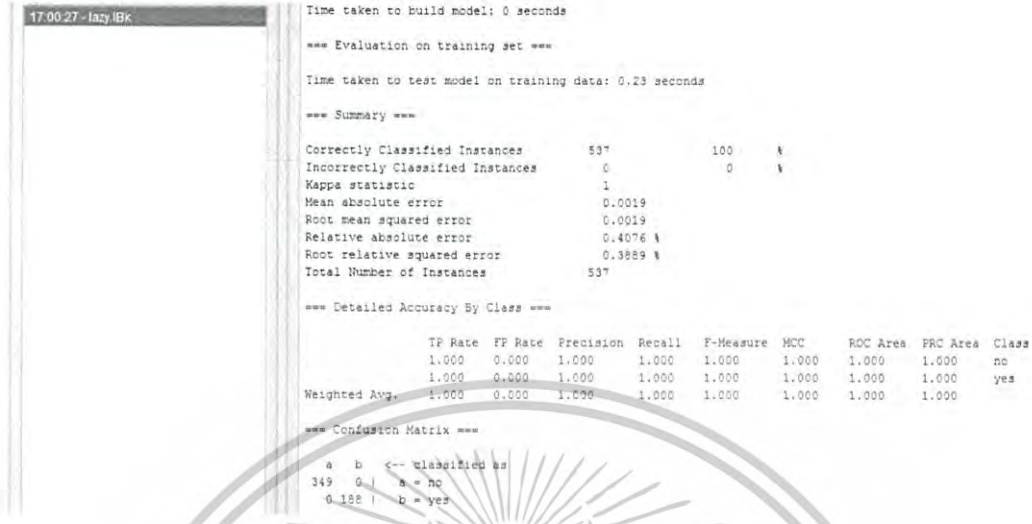


```
16:59:55 - bayes.NaiveBayes

Time taken to build model: 0.02 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.02 seconds

=== Summary ===

Correctly Classified Instances         415               77.2812 %
Incorrectly Classified Instances       122               22.7188 %
Kappa statistic                          0.4869
Mean absolute error                      0.2778
Root mean squared error                  0.4142
Relative absolute error                 61.0256 %
Root relative squared error             86.8297 %
Total Number of Instances              537

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.857    0.383    0.806      0.857    0.831      0.489    0.830     0.907     no
                 0.617    0.143    0.699      0.617    0.655      0.489    0.830     0.687     yes
Weighted Avg.    0.773    0.299    0.768      0.773    0.769      0.489    0.830     0.830

=== Confusion Matrix ===

   a   b   <-- classified as
 299  50 |   a = no
  72 116 |   b = yes
```

รูปที่ ข-6 ผลการวิเคราะห์ข้อมูลสำหรับการสร้างตัวแบบวิธีนาอีฟเบส์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2. วิธีเพื่อนบ้านใกล้สุด k ตัว



```
17:00:27 - lazy.IBk

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.23 seconds

=== Summary ===

Correctly Classified Instances         537               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0019
Root mean squared error                  0.0019
Relative absolute error                  0.4076 %
Root relative squared error              0.3889 %
Total Number of Instances              537

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000   1.000      1.000    1.000     1.000     no
                 1.000    0.000    1.000      1.000   1.000      1.000    1.000     1.000     yes
Weighted Avg.    1.000    0.000    1.000      1.000   1.000      1.000    1.000     1.000

=== Confusion Matrix ===

   a   b   <-- classified as
 349   0 |   a = no
   0 188 |   b = yes
```

รูปที่ ข-7 ผลการวิเคราะห์ข้อมูลสำหรับการสร้างตัวแบบวิธีเพื่อนบ้านใกล้สุด k ตัว

3. วิธีต้นไม้ตัดสินใจ

```
17:01:03 - trees.J48

Time taken to build model: 0.12 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances         467               86.9646 %
Incorrectly Classified Instances        70               13.0354 %
Kappa statistic                          0.7114
Mean absolute error                      0.2075
Root mean squared error                  0.3221
Relative absolute error                 45.5569 %
Root relative squared error             67.5304 %
Total Number of Instances              537

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.908    0.202    0.893      0.908   0.901      0.712    0.906     0.930     no
                 0.798    0.092    0.824      0.798   0.811      0.712    0.906     0.832     yes
Weighted Avg.    0.870    0.163    0.869      0.870   0.869      0.712    0.906     0.896

=== Confusion Matrix ===

   a   b   <-- classified as
 317  32 |   a = no
  38 150 |   b = yes
```

รูปที่ ข-8 ผลการวิเคราะห์ข้อมูลสำหรับการสร้างตัวแบบวิธีต้นไม้ตัดสินใจ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4. วิธีโครงข่ายประสาทเทียม

17:03:12 - functions.MultilayerPerceptron

```
Time taken to build model: 5.74 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances         420               78.2123 %
Incorrectly Classified Instances       117               21.7877 %
Kappa statistic                          0.5276
Mean absolute error                      0.2995
Root mean squared error                  0.3989
Relative absolute error                 65.7953 %
Root relative squared error             83.6277 %
Total Number of Instances              537

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.817    0.282    0.843      0.817    0.830      0.528    0.828     0.883     no
                 0.718    0.183    0.678      0.718    0.698      0.528    0.828     0.701     yes
Weighted Avg.    0.782    0.247    0.785      0.782    0.783      0.528    0.828     0.820

=== Confusion Matrix ===

   a   b   <-- classified as
 285  64 |   a = no
  53 135 |   b = yes
```

รูปที่ ข-9 ผลการวิเคราะห์ข้อมูลสำหรับการสร้างตัวแบบวิธีต้นไม้ตัดสินใจ

5. วิธีซัพพอร์ทเวกเตอร์แมชชีน

17:04:40 - functions.SMO

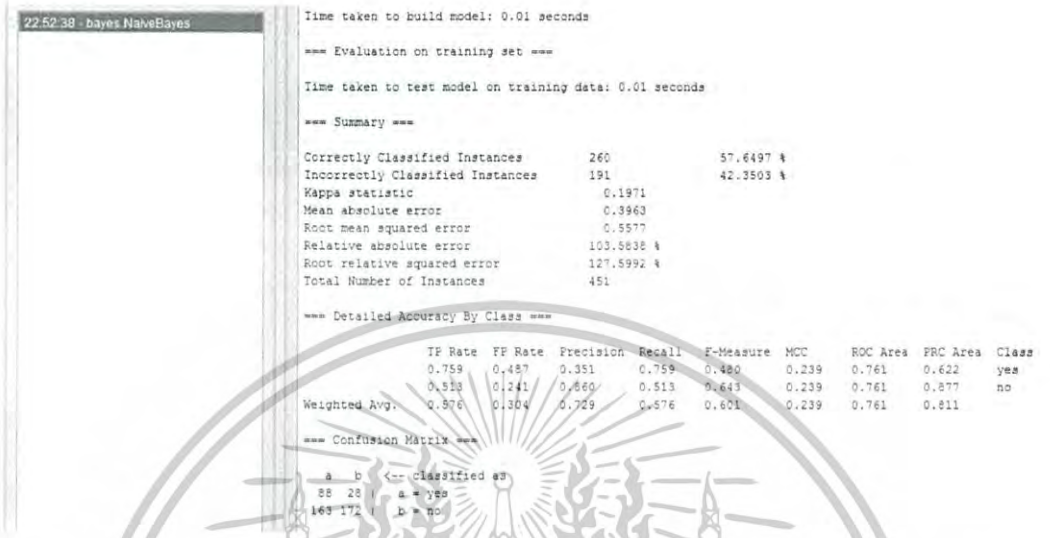
```
=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances         416               77.4674 %
Incorrectly Classified Instances       121               22.5326 %
Kappa statistic                          0.4648
Mean absolute error                      0.2253
Root mean squared error                  0.4747
Relative absolute error                 49.498  %
Root relative squared error             99.5149 %
Total Number of Instances              537

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.914    0.484    0.778      0.914    0.841      0.483    0.715     0.767     no
                 0.516    0.086    0.764      0.516    0.616      0.483    0.715     0.564     yes
Weighted Avg.    0.775    0.345    0.773      0.775    0.762      0.483    0.715     0.696

=== Confusion Matrix ===

   a   b   <-- classified as
 319  30 |   a = no
  91  97 |   b = yes
```

รูปที่ ข-10 ผลการวิเคราะห์ข้อมูลสำหรับการสร้างตัวแบบวิธีซัพพอร์ทเวกเตอร์แมชชีน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ข้อมูลการชำระเงินด้วยบัตรเครดิตของลูกค้า (Default Of Credit Card Clients)

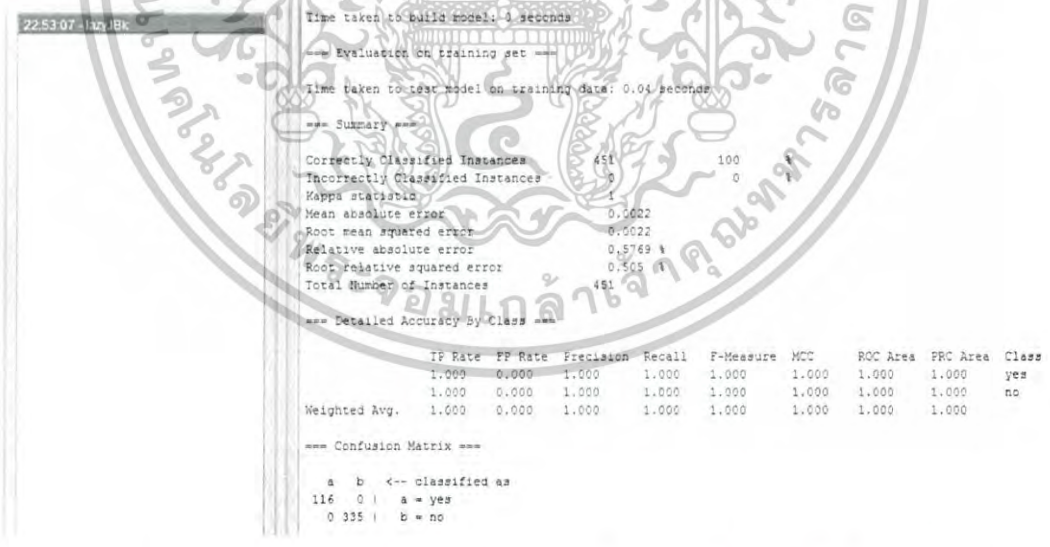
1. วิธีนาอีฟเบส์



```
22:52:38 - bayes.NaiveBayes

Time taken to build model: 0.01 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances         260               57.6497 %
Incorrectly Classified Instances       191               42.3503 %
Kappa statistic                          0.1971
Mean absolute error                      0.3963
Root mean squared error                  0.5577
Relative absolute error                103.5838 %
Root relative squared error            127.5992 %
Total Number of Instances              451

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.759     0.487     0.351     0.759     0.480     0.239     0.761     0.622     yes
               0.513     0.241     0.860     0.513     0.643     0.239     0.761     0.877     no
Weighted Avg.  0.576     0.304     0.729     0.576     0.601     0.239     0.761     0.811

=== Confusion Matrix ===

   a   b   <-- classified as
  88  28 |   a = yes
 163 172 |   b = no
```

รูปที่ ข-11 ผลการวิเคราะห์ข้อมูลสำหรับการสร้างตัวแบบวิธีนาอีฟเบส์

2. วิธีเพื่อนบ้านใกล้สุด k ตัว



```
22:53:07 - lazy.IBk

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.04 seconds

=== Summary ===

Correctly Classified Instances         451               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0022
Root mean squared error                  0.0022
Relative absolute error                  0.5769 %
Root relative squared error              0.505  %
Total Number of Instances              451

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               1.000     0.000     1.000     1.000     1.000     1.000     1.000     1.000     yes
               1.000     0.000     1.000     1.000     1.000     1.000     1.000     1.000     no
Weighted Avg.  1.000     0.000     1.000     1.000     1.000     1.000     1.000     1.000

=== Confusion Matrix ===

   a   b   <-- classified as
 116   0 |   a = yes
   0 335 |   b = no
```

รูปที่ ข-12 ผลการวิเคราะห์ข้อมูลสำหรับการสร้างตัวแบบวิธีเพื่อนบ้านใกล้สุด k ตัว

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3. วิธีต้นไม้ตัดสินใจ

22:53:35 - trees.J48

```
Time taken to build model: 0.01 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances         419               92.9047 %
Incorrectly Classified Instances        32                7.0953 %
Kappa statistic                          0.7996
Mean absolute error                      0.123
Root mean squared error                  0.248
Relative absolute error                 32.1411 %
Root relative squared error             56.7316 %
Total Number of Instances              451

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.750    0.009    0.967      0.750   0.845      0.810    0.926     0.869     yes
                 0.991    0.250    0.920      0.991   0.954      0.810    0.926     0.959     no
Weighted Avg.    0.929    0.188    0.932      0.929   0.926      0.810    0.926     0.936

=== Confusion Matrix ===

   a   b   <-- classified as
  87  29 |   a = yes
   3 332 |   b = no
```

รูปที่ ข-13 ผลการวิเคราะห์ข้อมูลสำหรับการสร้างตัวแบบวิธีต้นไม้ตัดสินใจ

4. วิธีโครงข่ายประสาทเทียม

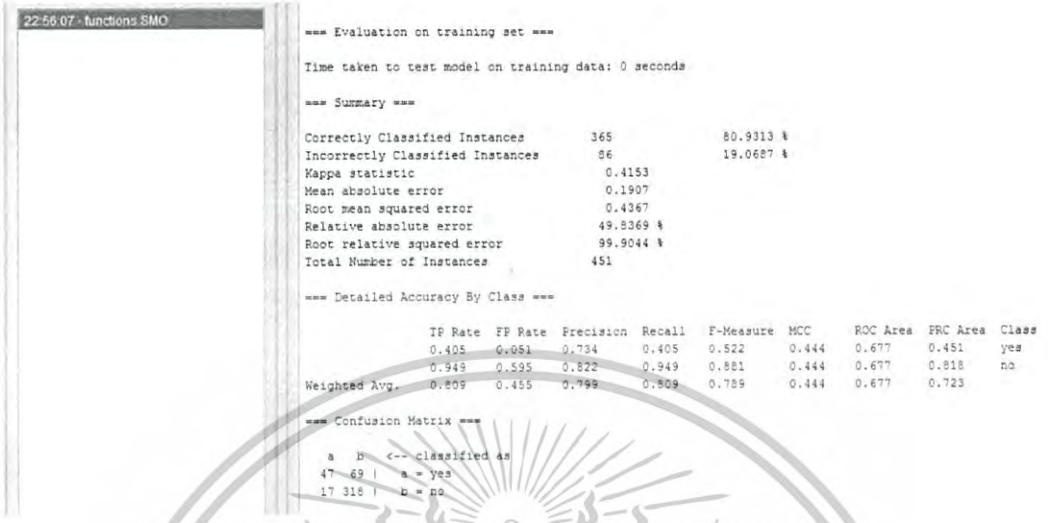
22:54:46 - functions.MultilayerPerceptron

```
Time taken to build model: 7.58 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances         389               86.2528 %
Incorrectly Classified Instances        62               13.7472 %
Kappa statistic                          0.6254
Mean absolute error                      0.2338
Root mean squared error                  0.3411
Relative absolute error                 61.0998 %
Root relative squared error             78.028  %
Total Number of Instances              451

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.672    0.072    0.765      0.672   0.716      0.628    0.791     0.596     yes
                 0.928    0.328    0.891      0.928   0.909      0.628    0.791     0.879     no
Weighted Avg.    0.863    0.262    0.859      0.863   0.860      0.628    0.791     0.806

=== Confusion Matrix ===

   a   b   <-- classified as
  78  38 |   a = yes
  24 311 |   b = no
```

รูปที่ ข-14 ผลการวิเคราะห์ข้อมูลสำหรับการสร้างตัวแบบวิธีต้นไม้ตัดสินใจ

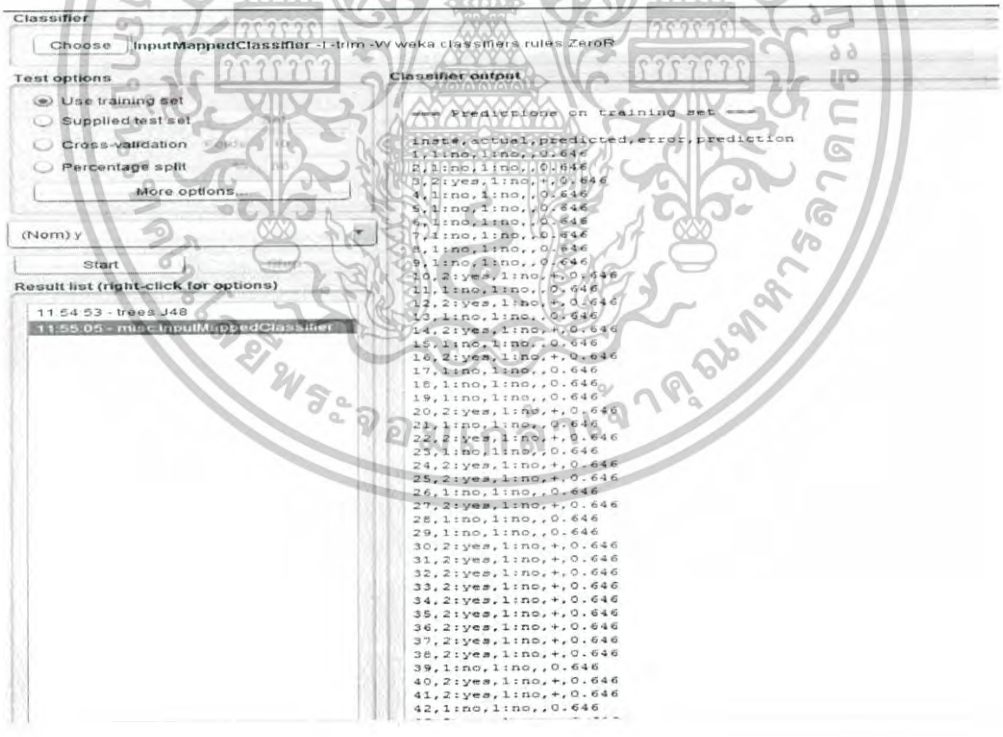
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

5. วิธีซัพพอร์ทเวกเตอร์แมชชีน



```
22:56:07 - functions.SMO

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances         365               80.9313 %
Incorrectly Classified Instances        86               19.0687 %
Kappa statistic                          0.4153
Mean absolute error                      0.1907
Root mean squared error                  0.4367
Relative absolute error                 49.8369 %
Root relative squared error             99.9044 %
Total Number of Instances              451

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.405    0.051    0.734      0.405   0.522      0.444    0.677     0.451     yes
               0.949    0.595    0.822      0.949   0.881      0.444    0.677     0.818     no
Weighted Avg.  0.809    0.455    0.799      0.809   0.789      0.444    0.677     0.723

=== Confusion Matrix ===

   a   b   <-- classified as
  47  69 |   a = yes
  17 318 |   b = no
```

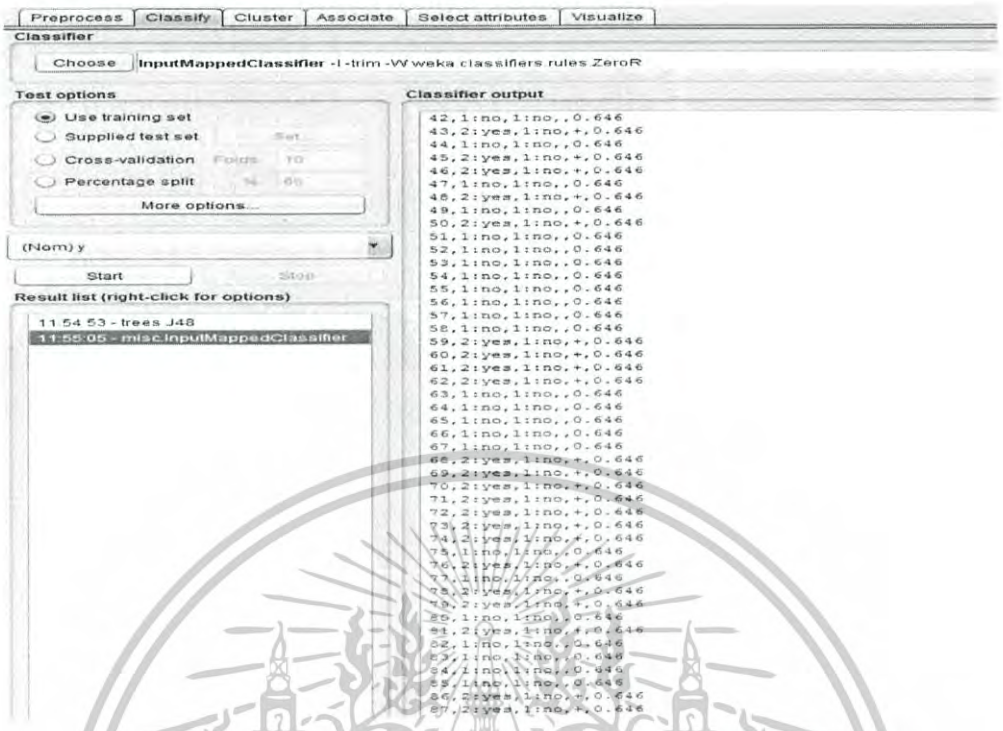
รูปที่ ข-15 ผลการวิเคราะห์ข้อมูลสำหรับการสร้างตัวแบบวิธีซัพพอร์ทเวกเตอร์แมชชีน



```
Classifier
Choose  InputMappedClassifier -I -trim -W weka.classifiers.rules.ZeroR

Test options                    Classifier output
 Use training set               === Predictions on training set ===
 Supplied test set              inst#,actual,predicted,error,prediction
 Cross-validation               1,1:no,1:no,,0.646
 Percentage split               2,1:no,1:no,,0.646
 More options...                3,2:yes,1:no,+,0.646
                                4,1:no,1:no,,0.646
(Nom) y                         5,1:no,1:no,,0.646
                                6,1:no,1:no,,0.646
Start                           7,1:no,1:no,,0.646
                                8,1:no,1:no,,0.646
Result list (right-click for options)   9,1:no,1:no,,0.646
                                10,2:yes,1:no,+,0.646
11:54:53 - trees.J48            11,1:no,1:no,,0.646
11:55:05 - misc.InputMappedClassifier   12,2:yes,1:no,+,0.646
                                13,1:no,1:no,,0.646
                                14,2:yes,1:no,+,0.646
                                15,1:no,1:no,,0.646
                                16,2:yes,1:no,+,0.646
                                17,1:no,1:no,,0.646
                                18,1:no,1:no,,0.646
                                19,1:no,1:no,,0.646
                                20,2:yes,1:no,+,0.646
                                21,1:no,1:no,,0.646
                                22,2:yes,1:no,+,0.646
                                23,1:no,1:no,,0.646
                                24,2:yes,1:no,+,0.646
                                25,2:yes,1:no,+,0.646
                                26,1:no,1:no,,0.646
                                27,2:yes,1:no,+,0.646
                                28,1:no,1:no,,0.646
                                29,1:no,1:no,,0.646
                                30,2:yes,1:no,+,0.646
                                31,2:yes,1:no,+,0.646
                                32,2:yes,1:no,+,0.646
                                33,2:yes,1:no,+,0.646
                                34,2:yes,1:no,+,0.646
                                35,2:yes,1:no,+,0.646
                                36,2:yes,1:no,+,0.646
                                37,2:yes,1:no,+,0.646
                                38,2:yes,1:no,+,0.646
                                39,1:no,1:no,,0.646
                                40,2:yes,1:no,+,0.646
                                41,2:yes,1:no,+,0.646
                                42,1:no,1:no,,0.646
```

รูปที่ ข-16 ผลการวิเคราะห์ข้อมูลสำหรับการทำนายตัวแบบด้วยวิธีนาอีฟเบส์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ ข-17 ผลการวิเคราะห์ข้อมูลสำหรับการทำนายตัวแบบด้วยวิธีนาอีฟเบส์ (ต่อ)

ตัวอย่างที่1 การคำนวณค่าความถูกต้อง (Accuracy) และค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) ของการวิเคราะห์ข้อมูลข้อมูลโรคมะเร็งเต้านมของรัฐวิสคอนซิน สำหรับการทำนายตัวแบบด้วยวิธีต้นไม้ตัดสินใจ

จากรูปที่ ข-3

$$\text{ค่าความถูกต้อง (Accuracy)} = \frac{TP+TN}{TP+TN+FP+FN} \times 100$$

$$= \frac{309+164}{309+164+9+7} \times 100$$

$$= 96.728\%$$

จากรูปที่ ข-16

1) กำหนดให้ค่า $y_i = 1$ ได้จากในกรณีที่ค่าจริง (actual) ใน class attribute ของข้อมูลโรคมะเร็งเต้านมในระเบียนนั้นตรงกับค่าทำนาย (predicted)

เช่น ระเบียนที่ 1 ค่าจริง (actual) = 1 : ไม่เป็นโรคมะเร็งเต้านม

ค่าทำนาย (predicted) = 1 : ไม่เป็นโรคมะเร็งเต้านม

จะได้ $y_i = y_1 = 1$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2) กำหนดให้ค่า $y_i = 0$ ได้จากในกรณีที่ค่าจริง (actual) ใน class attribute ของข้อมูลโรคมะเร็งเต้านมในระเบียนนั้นไม่ตรงกับค่าทำนาย (predicted)

เช่น ระเบียนที่ 3 ค่าจริง (actual) = 2 : เป็นโรคมะเร็งเต้านม

ค่าทำนาย (predicted) = 1 : ไม่เป็นโรคมะเร็งเต้านม

จะได้ $y_i = y_3 = 0$

กำหนดให้ค่า $\hat{y}_i$ ได้จาก ค่าทำนาย (prediction) ซึ่งอยู่ที่คอลัมน์ขวาสุดของระเบียนนั้นในช่อง Classifier output

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ลำดับที่	y_i	$\hat{y}_i$	e_i^2
1	0	0.968	0.937024
2	1	0.986	0.000196
3	1	0.968	0.001024
4	1	1	0
5	1	0.986	0.000196
6	1	0.986	0.000196
7	1	0.986	0.000196
8	1	0.986	0.000196
9	1	0.986	0.000196
10	0	0.84	0.7056
11	1	0.986	0.000196
12	1	0.968	0.001024
13	1	0.986	0.000196
14	1	0.968	0.001024
15	1	0.986	0.000196
16	1	1	0
17	1	0.986	0.000196
18	1	0.986	0.000196
19	1	0.986	0.000196
20	1	0.968	0.001024
21	1	0.986	0.000196
22	1	0.968	0.001024
23	0	0.968	0.937024
24	1	0.968	0.001024
25	1	0.968	0.001024
26	1	0.986	0.000196
27	1	0.968	0.001024
28	1	0.986	0.000196
29	1	0.986	0.000196
30	1	0.968	0.001024
31	1	0.8	0.04
32	1	1	0
33	1	0.968	0.001024
34	1	0.968	0.001024
35	0	0.986	0.972196
36	1	1	0
37	1	0.968	0.001024
38	1	1	0
39	1	0.986	0.000196

ลำดับที่	y_i	$\hat{y}_i$	e_i^2
40	1	0.968	0.001024
41	1	1	0
42	1	0.986	0.000196
43	0	0.84	0.7056
44	1	0.986	0.000196
45	1	0.917	0.006889
46	1	0.917	0.006889
47	1	0.986	0.000196
48	1	0.968	0.001024
49	1	0.84	0.0256
50	1	0.968	0.001024
51	1	0.986	0.000196
52	1	0.84	0.0256
53	1	0.986	0.000196
54	1	0.986	0.000196
55	1	0.986	0.000196
56	1	0.986	0.000196
57	1	0.986	0.000196
58	1	0.986	0.000196
59	1	0.968	0.001024
60	1	0.968	0.001024
61	1	0.917	0.006889
62	1	0.968	0.001024
63	1	0.986	0.000196
64	1	0.986	0.000196
65	1	0.986	0.000196
66	1	0.986	0.000196
67	1	0.986	0.000196
68	1	0.968	0.001024
69	1	1	0
70	1	0.8	0.04
71	0	0.986	0.972196
72	1	0.968	0.001024
73	1	0.968	0.001024
74	1	0.968	0.001024
75	1	0.986	0.000196
76	1	0.968	0.001024
77	1	0.84	0.0256
78	1	0.968	0.001024

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ลำดับที่	y_i	$\hat{y}_i$	e_i^2
79	1	0.968	0.001024
80	1	0.84	0.0256
81	1	0.968	0.001024
82	1	0.986	0.000196
83	1	0.986	0.000196
84	1	0.986	0.000196
85	1	0.986	0.000196
86	1	0.968	0.001024
87	1	0.917	0.006889
88	1	1	0
89	1	0.968	0.001024
90	1	0.986	0.000196
91	1	0.986	0.000196
92	1	0.986	0.000196
93	1	0.986	0.000196
94	1	0.986	0.000196
95	1	0.986	0.000196
96	1	0.986	0.000196
97	1	0.968	0.001024
98	1	0.986	0.000196
99	1	0.968	0.001024
100	1	0.986	0.000196
101	1	1	0
102	1	0.986	0.000196
103	1	0.986	0.000196
104	1	0.968	0.001024
105	1	0.986	0.000196
106	1	1	0
107	1	0.968	0.001024
108	1	0.968	0.001024
109	1	0.986	0.000196
110	1	0.986	0.000196
111	1	0.986	0.000196
112	1	0.986	0.000196
113	1	0.968	0.001024
114	1	0.968	0.001024
115	1	0.986	0.000196
116	1	0.986	0.000196
117	1	0.968	0.001024

ลำดับที่	y_i	$\hat{y}_i$	e_i^2
118	1	0.968	0.001024
119	1	0.986	0.000196
120	1	0.968	0.001024
121	1	0.917	0.006889
122	1	0.986	0.000196
123	1	0.986	0.000196
124	1	0.986	0.000196
125	1	0.968	0.001024
126	0	0.986	0.972196
127	1	0.968	0.001024
128	1	0.968	0.001024
129	1	0.986	0.000196
130	1	0.968	0.001024
131	1	0.986	0.000196
132	1	0.986	0.000196
133	1	0.986	0.000196
134	1	0.667	0.110889
135	1	0.986	0.000196
136	1	0.986	0.000196
137	1	0.968	0.001024
138	1	0.986	0.000196
139	1	0.986	0.000196
140	1	0.968	0.001024
141	1	0.968	0.001024
142	1	0.986	0.000196
143	1	0.986	0.000196
144	1	0.968	0.001024
145	1	0.986	0.000196
146	1	0.968	0.001024
147	1	0.968	0.001024
148	1	0.968	0.001024
149	1	0.986	0.000196
150	1	0.986	0.000196
151	1	0.968	0.001024
152	1	0.986	0.000196
153	1	1	0
154	1	0.968	0.001024
155	1	0.968	0.001024
156	1	0.986	0.000196

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ลำดับที่	y_i	$\hat{y}_i$	e_i^2
157	1	0.968	0.001024
158	1	0.986	0.000196
159	1	0.968	0.001024
160	1	0.968	0.001024
161	1	0.84	0.0256
162	1	0.917	0.006889
163	1	0.84	0.0256
164	1	0.986	0.000196
165	1	0.968	0.001024
166	1	0.968	0.001024
167	1	0.968	0.001024
168	1	0.917	0.006889
169	1	0.986	0.000196
170	1	0.986	0.000196
171	1	0.968	0.001024
172	1	0.917	0.006889
173	1	1	0
174	1	0.917	0.006889
175	0	0.917	0.840889
176	1	0.968	0.001024
177	1	0.968	0.001024
178	1	0.968	0.001024
179	1	0.986	0.000196
180	1	0.986	0.000196
181	1	0.986	0.000196
182	1	0.968	0.001024
183	1	0.968	0.001024
184	1	0.917	0.006889
185	1	0.968	0.001024
186	1	0.986	0.000196
187	1	0.917	0.006889
188	1	0.968	0.001024
189	1	0.917	0.006889
190	1	0.986	0.000196
191	0	0.986	0.972196
192	1	0.986	0.000196
193	1	0.986	0.000196
194	1	0.986	0.000196
195	1	0.986	0.000196

ลำดับที่	y_i	$\hat{y}_i$	e_i^2
196	1	0.986	0.000196
197	1	0.917	0.006889
198	1	0.917	0.006889
199	1	0.986	0.000196
200	1	1	0
201	1	0.986	0.000196
202	1	0.986	0.000196
203	1	0.917	0.006889
204	1	0.968	0.001024
205	1	0.84	0.0256
206	0	0.917	0.840889
207	1	0.986	0.000196
208	1	1	0
209	1	1	0
210	1	0.968	0.001024
211	1	0.986	0.000196
212	1	0.968	0.001024
213	1	0.986	0.000196
214	1	0.986	0.000196
215	1	0.968	0.001024
216	1	1	0
217	1	0.986	0.000196
218	1	0.986	0.000196
219	1	0.968	0.001024
220	1	0.986	0.000196
221	1	0.986	0.000196
222	0	0.968	0.937024
223	1	0.968	0.001024
224	1	0.968	0.001024
225	1	0.667	0.110889
226	1	0.986	0.000196
227	1	0.917	0.006889
228	1	0.986	0.000196
229	1	1	0
230	1	0.986	0.000196
231	1	0.917	0.006889
232	1	0.986	0.000196
233	1	0.986	0.000196
234	1	0.917	0.006889

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ลำดับที่	y_i	$\widehat{y_i}$	e_i^2
235	1	0.968	0.001024
236	1	0.986	0.000196
237	1	0.986	0.000196
238	1	0.917	0.006889
239	1	0.986	0.000196
240	1	0.968	0.001024
241	1	0.986	0.000196
242	1	0.986	0.000196
243	1	0.986	0.000196
244	1	0.8	0.04
245	1	1	0
246	1	0.986	0.000196
247	1	0.84	0.0256
248	1	0.968	0.001024
249	1	0.986	0.000196
250	1	0.986	0.000196
251	0	0.84	0.7056
252	1	0.968	0.001024
253	1	0.968	0.001024
254	1	0.8	0.04
255	1	0.84	0.0256
256	1	0.986	0.000196
257	1	0.968	0.001024
258	1	0.968	0.001024
259	1	0.986	0.000196
260	1	0.986	0.000196
261	1	0.986	0.000196
262	1	0.84	0.0256
263	1	0.986	0.000196
264	1	0.968	0.001024
265	1	0.986	0.000196
266	1	0.986	0.000196
267	1	0.986	0.000196
268	1	0.84	0.0256
269	1	0.968	0.001024
270	1	0.84	0.0256
271	1	0.986	0.000196
272	1	0.986	0.000196
273	1	0.968	0.001024

ลำดับที่	y_i	$\widehat{y_i}$	e_i^2
274	1	0.986	0.000196
275	1	0.986	0.000196
276	1	0.986	0.000196
277	1	0.986	0.000196
278	1	0.986	0.000196
279	1	0.986	0.000196
280	1	0.986	0.000196
281	1	0.968	0.001024
282	1	0.986	0.000196
283	1	0.84	0.0256
284	1	0.986	0.000196
285	1	0.986	0.000196
286	1	0.84	0.0256
287	1	0.986	0.000196
288	1	0.986	0.000196
289	1	0.986	0.000196
290	0	0.8	0.64
291	1	0.968	0.001024
292	1	0.84	0.0256
293	1	0.986	0.000196
294	1	0.84	0.0256
295	1	0.986	0.000196
296	1	0.968	0.001024
297	1	0.84	0.0256
298	1	0.968	0.001024
299	1	0.986	0.000196
300	1	0.986	0.000196
301	1	0.986	0.000196
302	1	0.986	0.000196
303	0	0.968	0.937024
304	1	0.968	0.001024
305	1	0.986	0.000196
306	1	0.986	0.000196
307	1	0.986	0.000196
308	1	0.917	0.006889
309	1	0.986	0.000196
310	1	0.986	0.000196
311	1	0.986	0.000196
312	1	0.986	0.000196

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ลำดับที่	y_i	$\hat{y_i}$	e_i^2
313	1	0.986	0.000196
314	1	0.986	0.000196
315	1	0.968	0.001024
316	1	0.986	0.000196
317	1	0.986	0.000196
318	1	0.84	0.0256
319	1	1	0
320	1	0.968	0.001024
321	1	1	0
322	1	0.986	0.000196
323	1	0.968	0.001024
324	1	0.968	0.001024
325	1	0.986	0.000196
326	1	0.986	0.000196
327	1	0.986	0.000196
328	1	0.986	0.000196
329	1	0.986	0.000196
330	1	0.986	0.000196
331	1	0.986	0.000196
332	1	0.986	0.000196
333	1	0.986	0.000196
334	1	0.986	0.000196
335	1	0.968	0.001024
336	1	0.968	0.001024
337	1	0.968	0.001024
338	1	0.986	0.000196
339	1	0.986	0.000196
340	1	0.968	0.001024
341	0	0.84	0.7056
342	1	0.986	0.000196
343	1	0.986	0.000196
344	1	0.968	0.001024
345	1	0.986	0.000196
346	1	0.986	0.000196
347	1	0.986	0.000196
348	1	0.986	0.000196
349	1	0.986	0.000196
350	1	0.986	0.000196
351	1	0.968	0.001024

ลำดับที่	y_i	$\hat{y_i}$	e_i^2
352	1	0.986	0.000196
353	1	0.986	0.000196
354	1	0.986	0.000196
355	1	0.986	0.000196
356	1	0.986	0.000196
357	1	0.968	0.001024
358	1	0.968	0.001024
359	1	0.986	0.000196
360	1	0.986	0.000196
361	1	0.968	0.001024
362	1	0.986	0.000196
363	1	0.986	0.000196
364	0	0.667	0.444889
365	1	0.968	0.001024
366	1	0.986	0.000196
367	1	0.986	0.000196
368	1	0.986	0.000196
369	1	0.986	0.000196
370	1	0.986	0.000196
371	1	0.986	0.000196
372	1	0.986	0.000196
373	1	0.986	0.000196
374	1	0.986	0.000196
375	1	0.986	0.000196
376	1	0.986	0.000196
377	1	0.986	0.000196
378	1	0.986	0.000196
379	1	0.986	0.000196
380	1	0.84	0.0256
381	1	0.986	0.000196
382	1	0.986	0.000196
383	1	0.986	0.000196
384	1	0.986	0.000196
385	1	0.986	0.000196
386	1	0.986	0.000196
387	1	0.986	0.000196
388	1	0.986	0.000196
389	1	0.986	0.000196
390	1	0.986	0.000196

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ลำดับที่	y_i	$\hat{y}_i$	e_i^2
391	1	0.986	0.000196
392	1	0.968	0.001024
393	1	0.986	0.000196
394	1	0.986	0.000196
395	1	0.968	0.001024
396	1	0.968	0.001024
397	1	0.986	0.000196
398	1	0.968	0.001024
399	1	0.986	0.000196
400	1	0.986	0.000196
401	1	0.986	0.000196
402	1	0.986	0.000196
403	1	0.986	0.000196
404	1	0.968	0.001024
405	1	0.968	0.001024
406	1	0.986	0.000196
407	1	0.986	0.000196
408	1	0.986	0.000196
409	1	0.968	0.001024
410	1	0.986	0.000196
411	1	0.968	0.001024
412	1	0.917	0.006889
413	1	0.986	0.000196
414	1	0.986	0.000196
415	1	0.986	0.000196
416	1	0.986	0.000196
417	1	0.986	0.000196
418	1	0.986	0.000196
419	1	0.986	0.000196
420	1	1	0
421	1	0.917	0.006889
422	1	0.986	0.000196
423	1	0.968	0.001024
424	1	0.986	0.000196
425	1	0.968	0.001024
426	1	0.986	0.000196
427	1	0.986	0.000196
428	1	0.986	0.000196
429	1	0.986	0.000196

ลำดับที่	y_i	$\hat{y}_i$	e_i^2
430	1	0.84	0.0256
431	1	0.986	0.000196
432	1	0.986	0.000196
433	1	0.968	0.001024
434	1	1	0
435	1	0.986	0.000196
436	1	0.986	0.000196
437	1	0.986	0.000196
438	1	0.986	0.000196
439	1	0.986	0.000196
440	1	0.968	0.001024
441	1	0.986	0.000196
442	1	0.986	0.000196
443	1	0.986	0.000196
444	1	0.986	0.000196
445	1	0.986	0.000196
446	1	0.986	0.000196
447	1	0.986	0.000196
448	1	0.986	0.000196
449	1	0.986	0.000196
450	1	0.986	0.000196
451	1	0.986	0.000196
452	1	0.986	0.000196
453	1	0.986	0.000196
454	1	0.986	0.000196
455	1	0.986	0.000196
456	1	0.986	0.000196
457	1	0.968	0.001024
458	1	0.986	0.000196
459	1	0.986	0.000196
460	1	0.986	0.000196
461	1	0.986	0.000196
462	1	0.986	0.000196
463	1	0.986	0.000196
464	1	0.986	0.000196
465	1	0.968	0.001024
466	1	0.968	0.001024
467	1	0.968	0.001024
468	1	0.986	0.000196

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ลำดับที่	y_i	$\hat{y}_i$	e_i^2
469	1	0.986	0.000196
470	1	0.84	0.0256
471	1	0.986	0.000196
472	1	0.986	0.000196
473	1	0.986	0.000196
474	1	0.968	0.001024
475	1	0.986	0.000196
476	1	0.986	0.000196
477	1	0.986	0.000196
478	1	0.986	0.000196
479	1	0.986	0.000196
480	1	0.986	0.000196
481	1	0.986	0.000196
482	1	0.986	0.000196
483	1	0.986	0.000196
484	1	0.968	0.001024
485	1	0.986	0.000196
486	1	0.986	0.000196
487	1	0.986	0.000196
488	1	0.968	0.001024
489	1	0.968	0.001024
รวม			14.476691

ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) $= \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n}$

$= \frac{14.476691}{489} = 0.0296$

รากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE) $= \sqrt{\text{MSE}} = \sqrt{0.0296} = 0.1721$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้