การเปรียบเทียบประสิทธิภาพและการทำนายผล
ในการจำแนกประเภทโรคความดันโลหิตสูง

An Efficiency Comparison and Prediction
of Classification Hypertension Data

ชลิดา ยอดนครจง
วรัชยา เดวาหมัด
สุธีรา จุขุนทด
สุปราณี เพ่งพิศ

ปัญหาพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรบัณฑิต สาขาวิชาสถิติประยุกต์
ภาควิชาสถิติ คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ปีการศึกษา 2557

An Efficiency Comparison and Prediction of Classification Hypertension Data

CHALIDA YODNAKHORNJONG

WARATCHAYA DEWAHMUD

SUTEERA JUKUNTOD

SUPRANEE PENGPIS

A SPECIAL PROJECT SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF BACHELOR OF SCIENCE
IN APPLIED STATISTICS
FACULTY OF SCIENCE
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG
ACADEMIC YEAR 2014

หัวข้อปัญหาพิเศษ การเปรียบเทียบประสิทธิภาพและการทำนายผลในการจำแนกประเภทโรคความดันโลหิตสูง
An Efficiency Comparison and Prediction of Classification Hypertension Data

ชื่อนักศึกษา

นางสาวชลิดา	ยอดนครจง	รหัสนักศึกษา 54050677
นางสาววรัชยา	เดวาหมัด	รหัสนักศึกษา 54050751
นางสาวสุธีรา	จุขุนทด	รหัสนักศึกษา 54050780
นางสาวสุปราณี	เพ่งพิศ	รหัสนักศึกษา 54050783

ปริญญา วิทยาศาสตรบัณฑิต (สถิติประยุกต์)
ภาควิชา สถิติ
ปีการศึกษา 2557
อาจารย์ที่ปรึกษา รองศาสตราจารย์สายชล สินสมบูรณ์ทอง

คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง อนุมัติให้ปัญหาพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา วิทยาศาสตรบัณฑิต (สถิติประยุกต์) ประจำปีการศึกษา 2557

คณะกรรมการสอบ	ลายมือชื่อ
รองศาสตราจารย์ สายชล สินสมบูรณ์ทอง กรรมการและอาจารย์ที่ปรึกษา	
ดร.ชานินทร์ ศรีสุวรรณนภา กรรมการ	
ดร.กนกกรรณ์ ลี้โรจนาประภา กรรมการ	

ลิขสิทธิ์ของคณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

หัวข้อปัญหาพิเศษ	การเปรียบเทียบประสิทธิภาพและการทำนายผลในการจำแนกประเภทโรคความดันโลหิตสูง		
ชื่อนักศึกษา	นางสาวชลิดา	ยอดนครจง	รหัสนักศึกษา 54050677
	นางสาววรัชยา	เดวาหมัด	รหัสนักศึกษา 54050751
	นางสาวสุธีรา	จุขุนทด	รหัสนักศึกษา 54050780
	นางสาวสุปราณี	เพ่งพิศ	รหัสนักศึกษา 54050783
ปริญญา	วิทยาศาสตรบัณฑิต (สถิติประยุกต์)		
ภาควิชา	สถิติ		
ปีการศึกษา	2557		
อาจารย์ที่ปรึกษา	รองศาสตราจารย์สายชล สินสมบูรณ์ทอง		

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพและการทำนายผลในการจำแนกประเภทโรคความดันโลหิตสูง โดยวิธีความใกล้เคียงกันมากที่สุด วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ วิธีโครงข่ายประสาทเทียม และวิธีซัพพอร์ตเวกเตอร์แมชชีน เพื่อประเมินประสิทธิภาพการจำแนกประเภทและประเมินผลการทำนาย โดยใช้ข้อมูลการเป็นโรคความดันโลหิตสูง โดยแบ่งข้อมูลเป็นชุดสร้างตัวแบบ ชุดทดสอบความถูกต้องของตัวแบบ และชุดทำนายตัวแบบ ในอัตราส่วน 70, 20 และ 10 ตามลำดับ

จากการเปรียบเทียบข้อมูลการเป็นโรคความดันโลหิตสูง วิธีการจำแนกประเภทที่มีประสิทธิภาพการจำแนกดีที่สุดสำหรับข้อมูลการเป็นโรคความดันโลหิตสูง โดยเปรียบเทียบจาก ค่าความถูกต้อง ทเอ ค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุล คือ วิธี โครงข่ายประสาทเทียม ส่วนวิธีการจำแนกประเภทที่มีผลการทำนายที่ดีที่สุดสำหรับ ข้อมูลการเป็นโรคความดันโลหิตสูง โดยเปรียบเทียบจากค่าความคลาดเคลื่อนกำลังสองเฉลี่ย คือ วิธีโครงข่ายประสาทเทียม

คำสำคัญ : การจำแนกประเภท วิธีความใกล้เคียงกันมากที่สุด วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ วิธีโครงข่ายประสาทเทียม และวิธีซัพพอร์ตเวกเตอร์แมชชีน

Title	An Efficiency Comparison of Classification and Prediction of Classification Hypertension Data		
Students	Chalida	Yodnakhornjong	ID 54050677
	Waratchaya	Dewahmud	ID 54050751
	Suteera	Jukuntod	ID 54050780
	Supranee	Pengpis	ID 54050783
Degree	Bachelor of Science (Applied Statistics)		
Department	Statistics		
Academic Year	2014		
Advisor	Assoc.Prof.Saichon Sinsomboonthong		

ABSTRACT

This objective of research is to compare efficiency and prediction of classification hypertension data result classification model of K-Nearest Neighbor, Decision Tree, Neural Network and Support Vector Machine by using hypertension data. Data sets are divided into training data, testing data and prediction data in the ratio 70, 20 and 10 respectively.

From a comparison of hypertension data, the best efficiency classification of hypertension data by comparing accuracy, precision, recall and F-Measure represents Neuron Network. The best classification to prediction by comparing mean square error represents Neuron Network.

KEYWORDS : Classification, K-Nearest Neighbor, Decision Tree, Neural Network and Support Vector Machine

กิตติกรรมประกาศ

ปัญญาพิเศษฉบับนี้สำเร็จลุล่วงไปได้ด้วยดีและมีความถูกต้องในเนื้อหา เนื่องด้วยได้รับความอนุเคราะห์จาก รศ.สายชล สินสมบูรณ์ทอง อาจารย์ที่ปรึกษา ที่ได้ให้ความรู้ ให้คำแนะนำ ให้ความช่วยเหลือ และตรวจแก้ในการทำปัญญาพิเศษจนสำเร็จลุล่วงไปได้ด้วยดี

ขอขอบพระคุณ ดร.ชานินทร์ ศรีสุวรรณนภา และ ดร.กนกกรรณ์ ลิ้โรจนาประภา ที่เป็นอาจารย์คณะกรรมการ ที่ให้คำปรึกษาเกี่ยวกับปัญหาพิเศษฉบับนี้ทั้งหมด สำหรับแนวคิด วิธีการ และข้อมูลทุกอย่างที่เป็นประโยชน์ในการวิเคราะห์วิธีการจำแนกประเภทและการทำนายผล

ขอขอบพระคุณอาจารย์ภาคสถิติทุกท่าน ที่ได้ประสิทธิ์ประสาทวิชาความรู้ พร้อมทั้งให้คำแนะนำและช่วยเหลือในเรื่องต่างๆตลอดมา

ขอขอบพระคุณ โรงพยาบาลสีคิ้ว จังหวัดนครราชสีมา ที่ได้ให้ข้อมูลในการทำปัญญาพิเศษฉบับนี้จนสำเร็จลุล่วงไปด้วยดี

สุดท้ายนี้ขอขอบพระครอบครัวของผู้จัดทำ ซึ่งสนับสนุนในด้านกำลังทรัพย์ และให้กำลังใจเสมอมาและขอบคุณเพื่อนๆทุกคน ที่ให้คำปรึกษา ช่วยเหลือในการทำงานมาโดยตลอด จนปัญญาพิเศษฉบับนี้จนสำเร็จลุล่วงไปด้วยดี

ชลิดา ยอดนครจง
วรัชยา เดวาหมัด
สุธีรา จุขุนทด
สุปราณี เพ่งพิศ

สารบัญ

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญตาราง(ต่อ)

สารบัญตาราง(ต่อ)

สารบัญรูป

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

โรคความดันโลหิตสูงคือภาวะที่มีระดับความดันโลหิตสูงเรื้อรัง มีค่าตั้งแต่ 140/90 มิลลิเมตรปรอทขึ้นไป ผู้คนจำนวนมากมีภาวะความดันโลหิตสูงโดยที่ไม่ทราบว่าตนเองมีภาวะนี้ เนื่องจากเป็นโรคที่ไม่ค่อยปรากฏอาการที่ชัดเจนในช่วงแรก แต่เมื่อปล่อยนานไปโดยไม่ได้รับการดูแลรักษา แรงดันในหลอดเลือดที่สูงจะไปทำลายผนังหลอดเลือดและอวัยวะที่สำคัญทั่วร่างกาย จึงเรียกโรคนี้กันว่า "เพชฌฆาตเงียบ" องค์การอนามัยโลก รายงานว่าทั่วโลกมีผู้ที่มีความดันโลหิตสูงมากถึงพันล้านคน ซึ่ง 2 ใน 3 ของจำนวนนี้อยู่ในประเทศกำลังพัฒนา โดยประชากรวัยผู้ใหญ่ทั่วโลก 1 คนใน 3 คนมีภาวะความดันโลหิตสูง และประชากรวัยผู้ใหญ่ในเขตเอเชียตะวันออกเฉียงใต้ก็พบว่ามี 1 คน ใน 3 คน ที่มีภาวะความดันโลหิตสูงเช่นกันและได้คาดการณ์ว่าในปี พ.ศ. 2568 (ค.ศ. 2025) ประชากรวัยผู้ใหญ่ทั่วทั้งโลกจะป่วยเป็นโรคความดันโลหิตสูง 1.56 พันล้านคน

ขณะเดียวกันสมาพันธ์ความดันโลหิตสูงโลก (World Hypertension League) ได้ประมาณการว่าปัจจุบันมีผู้ที่มีความดันโลหิตสูง 1 ใน 3 ทั่วโลก จึงคาดว่ามีจำนวนผู้เป็นความดันโลหิตสูง 1.8 พันล้านคน

โรคความดันโลหิตสูงเป็นหนึ่งในสาเหตุสำคัญของการเสียชีวิตก่อนวัยอันควร โดยในแต่ละปีประชากรวัยผู้ใหญ่ทั่วโลกเสียชีวิตจากโรคนี้ถึงเกือบ 8 ล้านคน ส่วนประชากรในแถบเอเชียตะวันออกเฉียงใต้มีผู้เสียชีวิตจากโรคความดันโลหิตสูงประมาณ 1.5 ล้านคน ซึ่งโรคความดันโลหิตสูงนี้ยังเป็นสาเหตุของการเสียชีวิตเกือบร้อยละ 50 ด้วยโรคอัมพฤกษ์ อัมพาต และโรคหัวใจ (Hypertension fact sheet, 2011)

จากข้อมูลการสำรวจสุขภาพประชาชนไทยโดยการตรวจร่างกาย ครั้งที่ 4 (พ.ศ. 2551-2552) (วิชัย เอกพลากร, 2553) พบว่า ประชากรไทยที่มีอายุ 15 ปี ขึ้นไป มีภาวะความดันโลหิตสูงเกือบ 11 ล้านคน และสิ่งที่น่าวิตกอย่างยิ่ง คือ ในจำนวนผู้ที่มีความดันโลหิตสูงร้อยละ 60 ในชายและ 40 ในหญิงไม่เคยได้รับการวินิจฉัยมาก่อน (ไม่รู้ตัวว่าเป็นความดันโลหิตสูง) ร้อยละ 8-9 ได้รับการวินิจฉัยแต่ไม่ได้รับการรักษา ส่งผลให้อาการทวีความรุนแรงขึ้นเพราะไม่ได้รับการรักษาและในกลุ่มของผู้ป่วยที่ได้รับการรักษา พบว่าจำนวนประมาณน้อยกว่า 1 ใน 4 ไม่สามารถควบคุมความดันโลหิตได้ตามเกณฑ์ซึ่งมีเพียง 1 ใน 4 ที่ได้รับการรักษาและควบคุมความดันโลหิตได้

จากข้อมูลการสำรวจพฤติกรรมเสี่ยงโรคไม่ติดต่อและการบาดเจ็บ ปีพ.ศ. 2553 (สถิติสาธารณสุข, 2550) โดยการสัมภาษณ์กลุ่มตัวอย่างอายุตั้งแต่ 15 – 74 ปี ของสำนักโรคไม่ติดต่อ กรมควบคุมโรค รายงานว่ามี 1 ใน 5 (ร้อยละ 22.2) ของประชากรอายุ 35 – 74 ปี ไม่ได้รับการตรวจความดันโลหิตจากแพทย์ พยาบาล เจ้าหน้าที่สาธารณสุขหรืออาสาอาสมัครสาธารณสุขประจำหมู่บ้าน ภายใน 1 ปีที่ผ่านมา และในเพศชายไม่ได้รับการตรวจวัดความดันโลหิต ร้อยละ 26.8 ส่วนเพศหญิง ไม่ได้รับการตรวจวัดความดันโลหิต ร้อยละ 18

รายงานผลการคัดกรองโรคเบาหวานและโรคความดันโลหิตสูง ปีงบประมาณ 2553 (วิชัย เอกพลากร, 2553) จากการคัดกรองในประชากรอายุ 15 ปีขึ้นไป จำนวน 23,349,952 คน พบว่าเป็นกลุ่มปกติ 15,908,677 คน กลุ่มเสี่ยงสูง 6,268,415 คน เป็นกลุ่มป่วยหรือสงสัยป่วยรายใหม่

1,172,860 คน (แยกเป็นระดับ 1 (140 ≤ ระดับความดันตัวบน < 160 และ/หรือ 90 ≤ ระดับความดันตัวล่าง < 100) จำนวน 946,252 คน ระดับ 2 (160 ≤ ระดับความดันตัวบน < 180 และ/หรือ 100 ≤ ระดับความดันตัวล่าง < 110) จำนวน 163,918 คน และระดับ 3 (ระดับความดันตัวบน ≥ 180 และ/หรือ ระดับความดันตัวล่าง ≥ 110) จำนวน 62,690 คน) และจากรายงานข้อมูลผู้ป่วยกลุ่มเสี่ยงความดันโลหิตสูงปีงบประมาณ 2555 ซึ่งป่วยความดันโลหิตสูงในปีงบประมาณ 2556 พบว่าจากกลุ่มเสี่ยง 8,525,803 คน ป่วยเป็นโรคความดันโลหิตสูง 64,115 คน คิดเป็นร้อยละ 0.75 และ จากกลุ่มปกติ 12,059,557 คน ป่วยเป็นโรคความดันโลหิตสูง 26,449 คน คิดเป็นร้อยละ 0.22 รวมป่วยเป็นโรคความดันโลหิตสูง ปี 2556 จำนวน 90,564 คน

จากที่กระทรวงสาธารณสุข ได้ทำการเปิดเผยข้อมูลจากทางองค์การอนามัยโลก ซึ่งพบว่าในปัจจุบันมีผู้ป่วยที่เป็นโรคความดันโลหิตสูงเกือบถึงพันล้านคน โดย 2 ใน 3 ของผู้ป่วยโรคความดันโลหิตสูงอยู่ในประเทศกำลังพัฒนาและพบในวัยผู้ใหญ่ในเขตเอเชียตะวันออกเฉียงใต้ ซึ่งรวมถึงประเทศไทยด้วย ดังนั้นการนำข้อมูลของผู้ป่วยมาเปรียบเทียบประสิทธิภาพจำแนกประเภทและการทำนายการเกิดโรคความดันโลหิตสูงโดยใช้ วิธีความใกล้เคียงกันมากที่สุด (K-Nearest Neighbor : KNN) วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ (Decision Tree) วิธีโครงข่ายประสาทเทียม (Neural Network) และวิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) จึงเป็นสิ่งที่น่าสนใจ เพราะจะทำให้สามารถทราบได้ว่าวิธีใดเป็นวิธีที่ดีที่สุดในการจำแนกประเภทและทำนายการเกิดโรคความดันโลหิตสูง

1.2 วัตถุประสงค์ของการศึกษา

1. เพื่อเปรียบเทียบประสิทธิภาพในการจำแนกประเภทการเกิดโรคความดันโลหิตสูง โดยใช้เทคนิควิธีความใกล้เคียงกันมากที่สุด วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ วิธีโครงข่ายประสาทเทียม และวิธีซัพพอร์ตเวกเตอร์แมชชีน

2. เพื่อทำนายผลการเกิดโรคความดันโลหิตสูง โดยวิธีความใกล้เคียงกันมากที่สุด วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ วิธีโครงข่ายประสาทเทียม และวิธีซัพพอร์ตเวกเตอร์แมชชีน

1.3 ขอบเขตการศึกษา

การวิจัยครั้งนี้มุ่งเน้นการเปรียบเทียบประสิทธิภาพในการจำแนกและทำนายผลในการจำแนกประเภทการเกิดโรคความดันโลหิตสูงโดยใช้เทคนิคการทำเหมืองข้อมูล ในการวิจัยครั้งนี้จะศึกษา ตัวแปรอิสระ ได้แก่ เพศ น้ำหนัก ส่วนสูง กรรมพันธุ์ ดัชนีมวลกาย เกณฑ์ดัชนีมวลกาย การสูบบุหรี่ การดื่มเครื่องดื่มที่มีแอลกอฮอล์ การออกกำลังกาย การรับประทานอาหาร รอบเอว และเกณฑ์ความดัน ตัวแปรตาม คือ การเป็นโรคความดันโลหิตสูง ข้อมูลที่นำมาใช้ในการศึกษาครั้งนี้เป็นข้อมูลทุติยภูมิจากการเก็บรวบรวมข้อมูลจำนวนผู้ป่วยที่มารับการรักษาที่โรงพยาบาลสีคิ้วจังหวัดนครราชสีมา จากเวชระเบียนของโรงพยาบาล

ประชากร คือ ผู้ป่วยที่มารับการรักษาที่โรงพยาบาลสีคิ้ว ตั้งแต่ปี พ.ศ. 2499-2557

ตัวอย่าง คือ ผู้ป่วยที่เข้ามารับการรักษา ในช่วงเดือนมกราคม-มีนาคม ปี พ.ศ. 2557

1.4 นิยามศัพท์เฉพาะ

ดัชนีมวลกาย (Body Mass Index) เป็นค่าดัชนีที่คำนวณจากน้ำหนักและส่วนสูง เพื่อใช้เปรียบเทียบความสมดุลระหว่างน้ำหนักตัวต่อความสูงของมนุษย์

$$BMI = \frac{WEIGHT}{HEIGHT^2}$$

-ดัชนีมวลกายน้อยกว่า 18.5 กก./ม^2 แสดงว่ามีน้ำหนักน้อยเกินไป ซึ่งอาจจะเกิดจากนักกีฬาที่ออกกำลังกายมากและได้รับสารอาหารไม่เพียงพอ วิธีแก้ไขต้องรับประทานอาหารที่มีคุณภาพและมีปริมาณพลังงานเพียงพอ และออกกำลังกายอย่างเหมาะสม

-ดัชนีมวลกายระหว่าง 18.5-22.9กก./ม^2 แสดงว่ามีน้ำหนักปกติและมีปริมาณไขมันอยู่ในเกณฑ์ปกติ มักจะเป็นโรคเบาหวาน ความดันโลหิตสูงต่ำกว่าผู้ที่อ้วนกว่านี้

-ดัชนีมวลกายอยู่ระหว่าง 23-24.9 กก./ม^2 แสดงว่าเริ่มจะมีน้ำหนักเกินหากมีกรรมพันธุ์เป็นโรคเบาหวานหรือไขมันในเลือดสูงต้องพยายามลดน้ำหนักให้ดัชนีมวลกายต่ำกว่า 23

-ดัชนีมวลกายอยู่ระหว่าง 25-29.9 กก./ม^2 แสดงว่าจัดว่าเป็นคนอ้วนระดับ 1 และหากมีเส้นรอบเอวมากกว่า 90 ซม. (ชาย) 80 ซม. (หญิง) จะมีโอกาสเกิดโรคความดัน เบาหวานสูง จำเป็นต้องควบคุมอาหารและออกกำลังกาย

-ดัชนีมวลกายมากกว่า 30 กก./ม^2 แสดงว่าจัดว่าอ้วนระดับ2 เสี่ยงต่อการเกิดโรคที่มากับความอ้วนหากมีเส้นรอบเอวมากกว่าเกณฑ์ปกติจะเสี่ยงต่อการเกิดโรคสูง ต้องควบคุมอาหารและออกกำลังกายอย่างจริงจัง (กองออกกำลังกาย เพื่อสุขภาพ กรมอนามัย กระทรวงสาธารณสุข, 2556)

อัลกอริทึม (Algorithm) หมายถึง วิธีการหรือกระบวนการทำงานใดงานหนึ่งที่สามารถแบ่งขั้นตอนออกเป็นขั้นตอนย่อยๆ ที่แน่นอน ซึ่งเมื่อทราบขั้นตอนการทำงานที่แน่นนอนแล้วก็จะนำอัลกอริทึมที่ได้นั้นมาวาดเป็น Flowchart จากนั้นจึงแปลง Flowchart เป็นภาษาระดับสูงที่คอมพิวเตอร์เข้าใจ (จุฬาลักษณ์ ถาไชยลา, 2554)

การทำเหมืองข้อมูล (Data Mining) คือ กระบวนการที่กระทำกับข้อมูลจำนวนมากเพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น ในปัจจุบันการทำเหมืองข้อมูลได้ถูกนำไปประยุกต์ใช้ในงานหลายประเภททั้งในด้านธุรกิจที่ช่วยในการตัดสินใจของผู้บริหาร ในด้านวิทยาศาสตร์และการแพทย์ รวมทั้งในด้านเศรษฐกิจและสังคม การทำเหมืองข้อมูลเปรียบเสมือนวิวัฒนาการหนึ่งในการจัดเก็บและตีความหมายข้อมูล จากเดิมที่มีการจัดเก็บข้อมูลอย่างง่ายๆ มาสู่การจัดเก็บในรูปฐานข้อมูลที่สามารถดึงข้อมูลสารสนเทศมาใช้จนถึงการทำเหมืองข้อมูลที่สามารถค้นพบความรู้ที่ซ่อนอยู่ในข้อมูล (อดุลย์ ยิ้มงาน, 2557)

โรคความดันโลหิตสูง (Hypertension) หมายถึง สภาวะผิดปกติที่บุคคลมีระดับความดันโลหิตสูงขึ้นกว่าระดับปกติของคนส่วนใหญ่ ถือว่าเป็นสภาวะที่ต้องควบคุม เนื่องจากความดันโลหิตสูงทำให้เกิดความเสียหายและการเสื่อมสภาพของหลอดเลือดแดงทั่วร่างกาย นำไปสู่ภาวะหลอดเลือดแดงแข็งและอุดตันหรือหลอดเลือดแตก โรคที่จะเกิดขึ้นจากความดันโลหิตที่สูงผิดปกติมีหลายโรค คือ โรคหลอดเลือดหัวใจหรือโรคหัวใจขาดเลือด โรคหลอดเลือดสมองหรือโรคอัมพาต โรคหัวใจวาย โรค

ไตวายเรื้อรัง โรคสมองเสื่อม การรักษาควบคุมความดันโลหิตสูงให้ลดลงเป็นปกติจะสามารถป้องกันโรคร้ายแรงต่างๆ ที่กล่าวถึงได้เป็นส่วนมาก

วิธีการวัดความดันโลหิตที่เป็นมาตรฐาน คือ การวัดโดยเครื่องวัดความดันโลหิตแบบปรอทที่พบเห็นตามห้องตรวจของโรงพยาบาลโดยทั่วไป วิธีการวัดความดันโดยใช้เครื่องแบบอื่นๆ เช่น เครื่องอัตโนมัติที่แสดงตัวเลขไม่ได้เป็นวิธีที่ดีกว่าวิธีใช้เครื่องวัดแบบปรอท ค่าความดันที่วัดได้จะออกมา 2 ค่า คือตัวเลขค่าสูงและค่าต่ำ ค่าสูง คือระดับความดันโลหิตขณะที่หัวใจบีบตัว ค่าต่ำ คือ ระดับความดันโลหิตขณะที่หัวใจคลายตัว ตัวเลขทั้งสองค่าจะรายงานเป็นมิลลิเมตรปรอท ระดับของความดันโลหิตที่ถือว่าสูงผิดปกติ คือ ค่าสูงตั้งแต่ 140 มิลลิเมตรปรอทขึ้นไปหรือค่าต่ำตั้งแต่ 90 มิลลิเมตรปรอทลงไป โดยระดับความดันทั้งสองค่านี้ยิ่งสูงมากก็ยิ่งจะมีโอกาสเกิดโรคแทรกซ้อนต่างๆ ได้มาก ตามลำดับ (พวงทอง ไกรพิบูลย์, 2550)

ความใกล้เคียงกันมากที่สุด (K-Nearest Neighbor : KNN) หมายถึง กระบวนการตัดสินใจของกลุ่ม สำหรับแทนเงื่อนไขหรือกรณีใหม่ โดยการตรวจสอบจำนวนบางจำนวน หรือเงื่อนไขที่เหมือนกันหรือใกล้เคียงกันมากที่สุด ใช้เวลาในการคำนวณสูงเพราะการคำนวณเป็นการเพิ่มขึ้นแบบแฟคทอเรียล (factorial) ตามจุดทั้งหมด (วาทินี นุ้ยเพียร และคณะ, 2553)

แผนภาพต้นไม้ (Decision Tree) หมายถึง กระบวนการในการจัดแบ่งข้อมูล โดยจะมีลักษณะการทำงานเหมือนโครงสร้างต้นไม้ โดยที่แต่ละโหนด (Node) จะแสดงคุณลักษณะ (Attribute) ที่ใช้ทดสอบข้อมูล แต่ละกิ่งจะแสดงผลในการทดสอบและโหนดใบ (Leaf Node) จะแสดงกลุ่ม (Class) ที่กำหนดไว้ ซึ่งแผนภาพต้นไม้ เพื่อการตัดสินใจนี้ง่ายต่อการเข้าใจและเป็นเทคนิคที่ค่อนข้างแพร่หลาย (ภัทร์พงศ์ พงศ์ภัทรกานต์, 2552)

โครงข่ายประสาท (Neural Network) หมายถึง ตัวแบบทางคณิตศาสตร์สำหรับประมวลผลสารสนเทศ ได้มาจากการศึกษาข่ายงานไฟฟ้า (Bioelectric Network) เกิดจากการเชื่อมต่อระหว่างเซลล์ประสาท เป็นการเลียนแบบวิธีการทำงานของสมองมนุษย์ ซึ่งเป็นรูปแบบการคำนวณที่ค่อนข้างซับซ้อน โดยนำข้อมูลต่างๆ ไปใช้ในการวิเคราะห์ ตีความ หรือคาดคะเน (วาทินี นุ้ยเพียร และคณะ, 2553)

วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) หมายถึง กระบวนการสมการที่ใช้ในการจำแนกค่าคุณลักษณะของสองกลุ่มที่วางตัวอยู่ในพื้นที่คุณลักษณะ (Feature Space) ออกจากกันโดยจะสร้างเส้นแบ่งที่เป็นเส้นตรงขึ้นมา เพื่อให้ทราบว่าเส้นตรงที่แบ่งสองกลุ่มออกจากกันนั้นเส้นตรงใดดีที่สุด (ภัทร์พงศ์ พงศ์ภัทรกานต์, 2552)

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. ทำให้เข้าใจโรคความดันโลหิตสูงซึ่งถือว่าเป็นโรคที่ก่อให้เกิดโรคอื่น ตามมาและพบได้จำนวนมาก

2. ทำให้ทราบประสิทธิภาพในการจำแนกประเภทการเกิดความดันโลหิตสูง โดยใช้วิธีความใกล้เคียงกันมากที่สุด วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ วิธีโครงข่ายประสาท และวิธีซัพพอร์ตเวกเตอร์แมชชีน

3. ทำให้ทราบผลการทำนายการเกิดโรคความดันโลหิตสูง โดยวิธีความใกล้เคียงกันมากที่สุด วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ วิธีโครงข่ายประสาท และวิธีซัพพอร์ตเวกเตอร์แมชชีน

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้ ผู้วิจัยได้นำเสนอเนื้อหาที่เน้นถึงทฤษฎีและงานวิจัยที่เกี่ยวข้องโดยมีรายละเอียดของเนื้อหาประกอบด้วยหัวข้อย่อย 9 หัวข้อ ดังนี้

2.1 การทำเหมืองข้อมูล (Data Mining)
2.2 วิธีการแบ่งประเภทข้อมูลของวิธีการจำแนกประเภท
2.3 วิธีความใกล้เคียงกันมากที่สุด (K-Nearest Neighbor : KNN)
2.4 วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ (Decision Tree)
2.5 วิธีโครงข่ายประสาทเทียม (Neural Network)
2.6 วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)
2.7 การเปรียบเทียบประสิทธิภาพของวิธีการจำแนกประเภท
2.8 การเปรียบเทียบผลการทำนายของวิธีการจำแนกประเภท
2.9 งานวิจัยที่เกี่ยวข้อง

2.1 การทำเหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูลคือการวิเคราะห์ข้อมูลเพื่อแยกประเภท จำแนกรูปแบบและความสัมพันธ์ของข้อมูลจากฐานข้อมูลที่มีขนาดใหญ่หรือคลังข้อมูล (รุจิรา ธรรมสมบัติ, 2554) โดยมีเทคนิคต่างๆหลายวิธี ซึ่งรูปแบบการทำเหมืองข้อมูลจากฐานข้อมูลนั้นได้รวบรวมความรู้จากหลายแขนงเข้าไว้ด้วยกันซึ่งประกอบด้วยการเรียนรู้ของเครื่องจักร (Machine Learning) รวมกับวิทยาศาสตร์สารสนเทศ (Information Science) สถิติ (Statistic) และระบบฐานข้อมูล (Database System) โดยทั่วไปแล้วเทคนิคที่นำมาใช้ส่วนใหญ่มี 5 ประเภท

2.1.1 เทคนิคการจำแนกประเภท (Classification)

เป็นเทคนิคการจำแนกกลุ่มข้อมูลด้วยคุณลักษณะต่างๆ ที่ได้มีการกำหนดไว้แล้ว เทคนิคประเภทนี้เหมาะกับการสร้างตัวแบบเพื่อการพยากรณ์ค่าข้อมูล (Predictive Modeling) ในอนาคตจากที่ได้จำแนกกลุ่มข้อมูลตัวอย่างไว้แล้ว ซึ่งในลักษณะดังกล่าวเรียกว่าการเรียนรู้แบบมีผู้สอน (Supervised Learning) เทคนิคการจำแนกประเภทเป็นกระบวนการสร้างตัวแบบเพื่อจัดข้อมูลให้อยู่ในกลุ่มที่กำหนด ตัวอย่างเช่น การแบ่งประเภทลูกค้าว่าเชื่อถือได้หรือไม่ ซึ่งเป็นการสร้างตัวแบบโดยการเรียนรู้จากข้อมูลที่ได้กำหนดไว้เรียบร้อยแล้ว

2.1.2 เทคนิคการค้นหาความสัมพันธ์ (Association Rule Discovery)

เป็นเทคนิคที่ใช้ในการค้นหาความสัมพันธ์ของฐานข้อมูลที่มีขนาดใหญ่ เพื่อที่จะทำการวิเคราะห์ข้อมูลและหาสิ่งที่ซ่อนอยู่ในข้อมูลนั้น เช่น การวิเคราะห์ข้อมูลการซื้อขายสินค้าในซุปเปอร์มาร์เก็ต เพื่อทำการวางแผนการส่งเสริมการขาย (Promotion) และการเตรียมการวางแผนการเรียงชั้นวางสินค้า (Shelf) เช่น การวางน้ำอัดลมกับข้าวโพดคั่วไว้ใกล้กัน

2.1.3 เทคนิคการจัดกลุ่ม (Clustering)

เป็นเทคนิคการลดขนาดของข้อมูลด้วยการรวมกลุ่มตัวแปรที่มีลักษณะเดียวกันไว้ด้วยกัน ทำให้สามารถค้นหาข้อมูลที่ถูกละเลยไปได้ เทคนิคนี้มักถูกใช้เป็นขั้นตอนเบื้องต้นในการทำเหมืองข้อมูล และเหมาะกับข้อมูลที่ยังไม่มีกลุ่มอย่างชัดเจน จึงทำการรวมกลุ่มเพื่อหากลุ่มต่างๆ ของข้อมูล โดยจำนวนกลุ่มของข้อมูลแทนด้วย k กลุ่ม

2.1.4 เทคนิคการหาค่าที่แตกต่างจากค่ามาตรฐาน (Deviation Detection)

เป็นเทคนิควิธีในการหาค่าที่แตกต่างไปจากค่ามาตรฐาน หรือค่าที่คาดคิดไว้ว่าต่างไปมากน้อยเพียงใด โดยทั่วไปมักใช้วิธีทางสถิติ หรือการแสดงให้เห็นภาพ สำหรับเทคนิคนี้ใช้ในการตรวจสอบลายเซ็นต์ หรือปลอมบัตรเครดิต เป็นต้น

2.1.5 เทคนิคการวิเคราะห์ลำดับ (Sequential Analysis)

เป็นเทคนิคในการวิเคราะห์ลำดับเพื่อค้นหารูปแบบการปรากฏของข้อมูล ซึ่งปรากฏในรายการที่แยกออกมา เช่น ถ้าผู้ซื้อสินค้า A แล้วเขาจะซื้อสินค้า B ในภายหลัง เทคนิคนี้จะแตกต่างจากเทคนิคการค้นหาความสัมพันธ์ เพราะคำนึงถึงลำดับการซื้อด้วย

2.2 วิธีการแบ่งประเภทข้อมูลของวิธีการจำแนกประเภท (Classification)

การแบ่งประเภทข้อมูลคือกระบวนการสร้างตัวแบบเพื่อจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนด (พยูน พานิชย์กุล, 2548) เป็นการสร้างตัวแบบการจัดหมวดหมู่ได้จากกลุ่มตัวอย่างของข้อมูลที่ได้กำหนดไว้ล่วงหน้า และสามารถพยากรณ์กลุ่มของข้อมูลที่ยังไม่เคยนำมาจัดหมวดหมู่ได้ ตัวแบบที่ได้อาจอยู่ในรูปแบบแผนภาพต้นไม้เพื่อการตัดสินใจ (Decision Tree) หรือโครงข่ายประสาทเทียม (Neural Network)

ในการจัดหมวดหมู่จำเป็นต้องมีกลุ่มข้อมูลสำหรับการเรียนรู้ (Training Data) เพื่อให้ข้อมูลเรียนรู้และสร้างตัวแบบ (Model Construction) และทดสอบโดยกลุ่มข้อมูลสำหรับการทดสอบ (Testing Data) เพื่อประเมินความถูกต้องของตัวแบบ (Model Evaluation) อีกทั้งใช้ชุดข้อมูลที่ไม่เคยเห็นมาก่อน (Unseen Data) เพื่อทำการกำหนดกลุ่มให้กับข้อมูลใหม่ที่ได้มาหรือทำนายค่าออกมาตามที่ต้องการ เช่น การจัดหมวดหมู่ของผู้ยื่นบัตรเครดิต (Credits) เป็นระดับต่ำ ระดับกลาง และระดับสูง ของความเสี่ยงที่จะได้รับ หรือการอนุมัติบุคคลเข้าทำงานในลักษณะงานต่างๆ

2.3 วิธีความใกล้เคียงกันมากที่สุด (K-Nearest Neighbor:KNN)

วิธีความใกล้เคียงกันมากที่สุด เป็นวิธีการที่ได้รับความนิยมในการใช้งานเป็นอย่างมาก สาเหตุเนื่องจากเป็นวิธีการที่ง่ายและมีประสิทธิภาพซึ่งสามารถนำไปประยุกต์ใช้กับงานได้อย่างหลากหลาย (Troyanskaya, O. et al., 2001) เช่น งานทางด้านการจำแนกข้อมูล (Classification) รวมถึงงานทางด้านการแทนที่ข้อมูลที่สูญหาย (Missing Values Imputation) ซึ่งมีวิธีการดำเนินการดังนี้

2.3.1 กำหนดค่า k เพื่อใช้พิจารณาสมาชิกที่อยู่ใกล้กันมากที่สุด เช่น k = 3 คือจะพิจารณาเฉพาะข้อมูล 3 ตัวแรกที่อยู่ใกล้กับจุดที่ต้องการจะทำนาย

2.3.2 คำนวณหาระยะห่างระหว่างข้อมูลตัวอย่างที่สนใจกับข้อมูลอื่นๆ ทุกตัวด้วยวิธี**ระยะห่างยูคลิเดียน** (Euclidian distance) จากสมการที่ 2-1

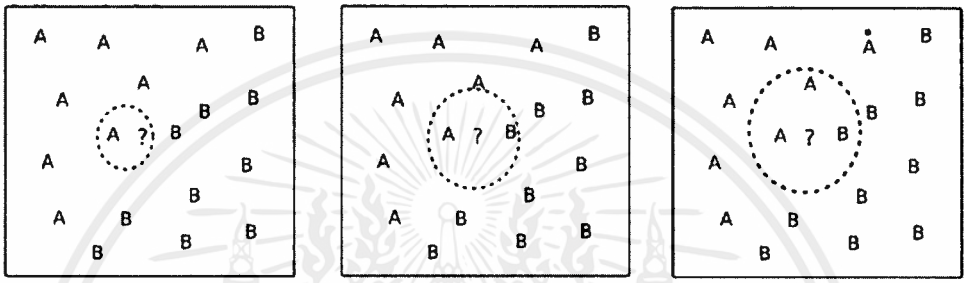
$$dist(x_i, x_j) = \sqrt{\sum_{k=1}^{n}(x_{i,k} - x_{j,k})^2} \qquad (2\text{-}1)$$

โดยที่ $dist(x_i, x_j)$ คือ ระยะห่างระหว่างตัวอย่าง x_i กับตัวอย่าง x_j

n คือ จำนวนคุณสมบัติทั้งหมดของตัวอย่าง

$x_{i,k}$ คือ คุณสมบัติที่ k ของตัวอย่าง x_i

2.3.3เลือกค่าข้อมูลที่มีค่าระยะห่างน้อยที่สุด k ตัวเพื่อนำมาพิจารณาหาคำตอบ ดังรูปที่ 2-1



(a) ความใกล้เคียงกันมากที่สุดโดยพิจารณาจากข้อมูล 1 ตัว
(b) ความใกล้เคียงกันมากที่สุดโดยพิจารณาจากข้อมูล 2 ตัว
(c) ความใกล้เคียงกันมากที่สุดโดยพิจารณาจากข้อมูล 3 ตัว

รูปที่ 2-1 ตัวอย่างของ K-Nearest Neighbor

2.4 วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ (Decision Tree)

เป็นตัวแบบทางคณิตศาสตร์เพื่อหาทางเลือกที่ดีที่สุด โดยการนำข้อมูลมาสร้างตัวแบบการพยากรณ์ในรูปแบบของโครงสร้างต้นไม้ซึ่งมีการเรียนรู้ข้อมูลแบบมีผู้สอน (Supervised Learning) สามารถสร้างตัวแบบการจัดกลุ่ม (Clustering) ได้จากกลุ่มตัวอย่างของข้อมูลฝึกหัด (Training Data Set) ได้โดยอัตโนมัติและสามารถพยากรณ์กลุ่มของรายการที่ยังไม่เคยนำมาจัดกลุ่มได้อีกด้วย (รุจิรา ธรรมสมบัติ, 2554)

โดยปกติมักประกอบด้วยกฎในรูปแบบ"ถ้า เงื่อนไข แล้ว ผลลัพธ์" เช่น

ถ้ารายได้สูงและยังไม่ได้แต่งงาน แล้วฐานะยากจน

"IF Income = High and Married = No THEN Risk = Poor"

ถ้ารายได้สูงและแต่งงานแล้ว แล้วฐานะร่ำรวย

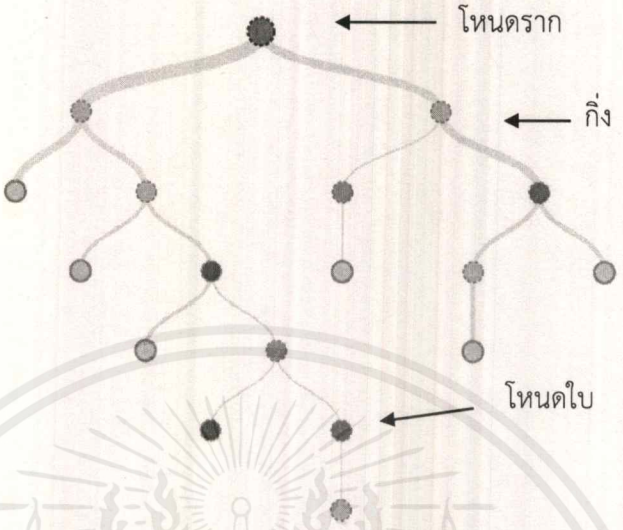
"IF Income = High and Married = Yes THEN Risk = Good"

2.4.1 ส่วนประกอบของต้นไม้เพื่อการตัดสินใจ

1) โหนด (Node) คือ คุณสมบัติต่างๆเป็นจุดที่แยกข้อมูลว่าจะให้ไปในทิศทางใดซึ่งโหนดที่อยู่สูงสุดเรียกว่า โหนดราก (Root Node)

2) กิ่ง (Branch) คือ คุณสมบัติของโหนดที่แตกออกมา โดยจำนวนของกิ่งจะเท่ากับคุณสมบัติของโหนด

3) ใบ (Leaf) คือ กลุ่มของผลลัพธ์ในการแยกแยะข้อมูล ซึ่งโหนดที่อยู่ล่างสุดเรียกว่า โหนดใบ (Leaf Node) โดยสามารถแสดงส่วนประกอบของแผนภาพต้นไม้เพื่อการตัดสินใจดังรูป 2-2



รูปที่ 2-2 ส่วนประกอบของแผนภาพต้นไม้เพื่อการตัดสินใจ

2.4.2 การสร้างแผนภาพต้นไม้เพื่อการตัดสินใจ

หลักการพื้นฐานของการสร้างแผนภาพต้นไม้เพื่อการตัดสินใจเป็นการสร้างในลักษณะจากบนลงล่าง (Top-Down) คือเริ่มจากการสร้างรากของต้นไม้ก่อนแล้วจึงแตกกิ่งไปจนถึงใบโดยแสดงขั้นตอนการสร้างแผนภาพต้นไม้เพื่อการตัดสินใจได้ดังนี้

1) ต้นไม้เริ่มต้นโดยมีโหนดเพียงโหนดเดียวแสดงถึงชุดข้อมูลฝึกหัด (Training Data Set)

2) ถ้าข้อมูลทั้งหมดอยู่ในกลุ่มเดียวกันแล้วให้โหนดนั้นเป็นใบและตั้งชื่อแยกตามกลุ่มของข้อมูลนั้น

3) ถ้าในโหนดมีข้อมูลหลายกลุ่มปะปนอยู่ จะต้องวัดค่าผลกำไร (Gain) ของแต่ละคุณลักษณะ (Attribute) เพื่อที่จะใช้เป็นเกณฑ์ในการคัดเลือกคุณลักษณะที่มีความสามารถในการแบ่งแยกข้อมูลออกเป็นกลุ่มต่างๆ ได้ดีที่สุด โดยคุณลักษณะที่มีผลกำไรมากที่สุดจะถูกเลือกให้เป็นตัวทดสอบหรือคุณลักษณะที่ใช้ในการตัดสินใจ โดยแสดงในรูปของโหนดบนต้นไม้

4) กิ่งของต้นไม้ ถูกสร้างขึ้นจากค่าต่างๆ ที่เป็นไปได้ของโหนดทดสอบ และข้อมูลจะถูกทดสอบ และข้อมูลจะถูกแบ่งออกตามกิ่งต่างๆ ที่สร้างขึ้น

5) ทำการวนซ้ำเพื่อหาคุณลักษณะที่มีผลกำไรมากที่สุด สำหรับข้อมูลที่ถูกแบ่งแยกออกมาในแต่ละกิ่งเพื่อนำคุณลักษณะนี้มาสร้างเป็นโหนดตัดสินใจต่อไป โดยที่คุณลักษณะที่ถูกเลือกมาเป็นโหนดแล้ว จะไม่ถูกเลือกมาอีกสำหรับโหนดในระดับต่างๆต่อไป

6) ทำการวนซ้ำเพื่อแบ่งข้อมูลและแตกกิ่งต้นไม้ไปเรื่อยๆ โดยการวนซ้ำจะสิ้นสุดก็ต่อเมื่อเงื่อนไขข้อใดข้อหนึ่งข้างบนนี้เป็นจริง

2.4.3 การคำนวณค่าเกนความรู้ (Information Gain)

แผนภาพต้นไม้เพื่อการตัดสินใจเป็นโครงสร้างที่ใช้แสดงกฎที่ได้จากเทคนิคการจำแนกประเภทข้อมูล โดยต้นไม้ตัดสินใจจะมีลักษณะคล้ายโครงสร้างต้นไม้ โดยที่แต่ละโหนดแสดงคุณลักษณะ (Attribute) ในการสร้างแผนภาพต้นไม้เพื่อการตัดสินใจ ปัญหาสำคัญที่ต้องพิจารณาคือควรจะตัดสินใจเลือกคุณลักษณะใดมาทำหน้าที่เป็นโหนดราก ในแต่ละขั้นตอนของการสร้างต้นไม้และต้นไม้ย่อย (Subtree) ของแผนภาพต้นไม้เพื่อการตัดสินใจ เกณฑ์ที่ใช้ช่วยประกอบการเลือกคุณลักษณะคือการคำนวณค่าเกนมาตรฐาน (Gain Criterion) ซึ่งเป็นค่าที่บ่งคุณลักษณะนั้นสามารถจำแนกกลุ่มของข้อมูลได้ดีเพียงใด โดยทดลองเลือกแต่ละคุณลักษณะที่เป็นไปได้จากชุดข้อมูลมาทำหน้าที่เป็นโหนดราก ถ้าคุณลักษณะใดให้เกนความรู้สูงที่สุด แสดงว่าคุณลักษณะนั้นสามารถจำแนกกลุ่มของข้อมูลได้ดีที่สุด การใช้ค่าผลกำไรสารสนเทศจะช่วยลดจำนวนครั้งของการทดสอบในการแยกแยะข้อมูล อีกทั้งยังรับประกันว่าแผนภาพต้นไม้เพื่อการตัดสินใจที่ได้ไม่มีความซับซ้อนมากจนเกินไป ซึ่งค่าผลกำไรสารสนเทศนั้นสามารถคำนวณได้จากสมการที่ 2-2

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

เมื่อ s แทนเซตของข้อมูลซึ่งประกอบด้วยข้อมูล s ระเบียน (record)
n แทนจำนวนกลุ่มทั้งหมดที่ต่างกันของข้อมูลชุดนั้น
s_i แทนจำนวนข้อมูลที่เป็นสมาชิกของ s และอยู่ในกลุ่มของ c_i
c_i แทนกลุ่มในลำดับที่ i มีค่าระหว่าง 1 ถึง n

ค่าเอ็นโทรปี (Entropy) ของคุณลักษณะ A $(E(A))$ซึ่งมีค่าของคุณลักษณะเป็น $(a_1, a_2, a_3, ..., a_v)$ หาได้ในสมการที่ 2-3

$$E(A) = \sum_{j=1}^{v} \frac{S_{ij} + ... + S_{nj}}{S} I(S_{1j}, S_{2j}, ..., S_{nj}) \qquad (2\text{-}3)$$

S_{ij} แทนข้อมูลจำนวนที่เป็นสมาชิกของ s ในกลุ่ม C_i จากการแบ่งข้อมูลด้วยค่าที่เป็นไปได้ของคุณลักษณะ A

ดังนั้นจะสามารถพิจารณาค่ามาตรฐานผลกำไรได้ดังสมการที่ 2-4

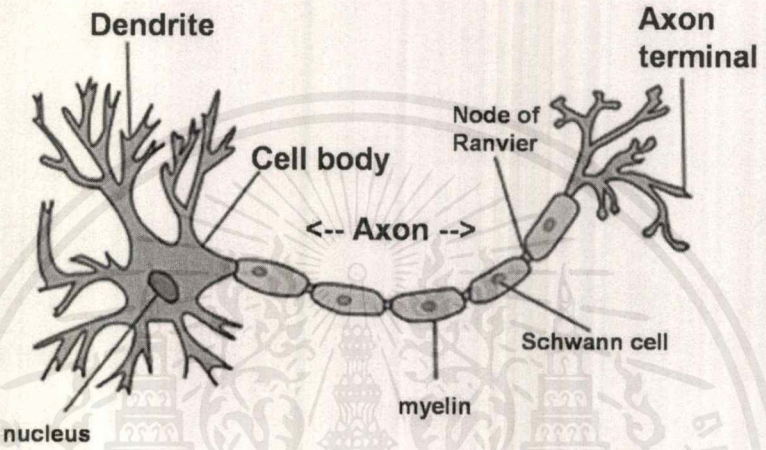
$$Gain(A) = I(S_{1j}, S_{2j}, ..., S_{nj}) - E(A) \qquad (2\text{-}4)$$

2.5 วิธีโครงข่ายประสาทเทียม (Neural Network)

โครงข่ายประสาทเทียมเป็นศาสตร์ที่จำลองแบบความสามารถของมนุษย์ด้านการเรียนรู้ จดจำและจำแนกสิ่งต่างๆ ซึ่งใช้สมองเป็นส่วนสำคัญ ในการประมวลระบบของโครงข่ายประสาทเทียมนั้นจะเลียนแบบการทำงานของระบบสมอง (ศุภกฤษ์ ชูธงชัย, 2546) คือ มีการส่งผ่านข้อมูลระหว่างกันโดยมีการเชื่อมต่อของเซลประสาท (Neuron) กันเป็นโครงข่ายร่างแหจำนวนมากและมีการประมวลผลในลักษณะขนาน (Parallel processing) สาเหตุหลักที่โครงข่ายประสาทเทียมเป็นที่นิยม

กันมากขึ้นเนื่องจากมีความยืดหยุ่นในการทำงานสูงและสามารถปรับตัวเองให้ทำงานในสภาพที่เปลี่ยนแปลง อีกทั้งไม่จำเป็นต้องทราบตัวแบบทางคณิตศาสตร์ (Mathematical model) ที่แน่นอนของกระบวนการ เพียงแต่ใช้ชุดข้อมูลที่ประกอบด้วยข้อมูลนำเข้า (Input data) และข้อมูลเป้าหมาย (Target data) ของกระบวนการในจำนวนที่มากพอมาใช้ในการสอน (Training) โครงข่ายประสาทเทียม

2.5.1 ความรู้พื้นฐานของระบบประสาท (Neural System Knowledge)



รูปที่ 2-3 โครงข่ายของเซลล์ประสาท

ภายในสมองของมนุษย์ประกอบด้วยหน่วยประมวลผลขนาดเล็กที่ เรียกว่า เซลล์ประสาท ซึ่งจะมีประมาณ 10 หน่วย ในเซลล์ประสาทแต่ละหน่วยดังแสดงใน รูปที่ 2-3 ประกอบด้วย ใยประสาท (Dendrites) ตัวเซลล์ (Cell body) และเส้นใยประสาท (Axon) ซึ่งแบ่งออกเป็น 4 บริเวณด้วยกันคือ

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

2) บริเวณรวมกระแสประสาท (Integration region) เป็นบริเวณที่มีการรวมกระแสประสาทก่อนที่จะเข้าสู่บริเวณการนำกระแสประสาทรวมออกจากเซลล์

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

4) บริเวณนำกระแสประสาทออก (Output region) เป็นบริเวณส่วนปลายของเส้นใยประสาทที่มีการแตกแขนงใช้ในการถ่ายทอดกระแสประสาทข้ามเซลล์ไปยังเซลล์ประสาทอื่น โดยผ่านทางใยประสาทของเซลล์ประสาทนั้น

2.5.2 การเรียนรู้ของโครงข่ายประสาทเทียม (Neural network learning)

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

1) การเรียนรู้แบบมีผู้สอน (Supervised learning)

การเรียนรู้แบบมีผู้สอนจะกำหนดข้อมูลฝึกหัด (Training data set) ให้กับโครงข่าย ซึ่งกลุ่มนี้ประกอบด้วยข้อมูลนำเข้า (Input data) และข้อมูลเป้าหมาย (Target data) ที่ต้องการ จากนั้นโครงข่ายจะทำการคำนวณค่าถ่วงน้ำหนักที่เหมาะสมให้กับข้อมูลฝึกหัด โดยคำตอบที่ได้จากโครงข่ายจะถูกคำนวณค่าความผิดพลาด(Error value) ว่ามีความห่างจากคำตอบที่ต้องการของข้อมูลนำเข้าในชุดเดียวกันมากน้อยเพียงใด ถ้ายังมีความผิดพลาดสูงอยู่ การฝึกหัดจะดำเนินต่อจนกว่าค่าความผิดพลาดจะลดลงต่ำกว่าค่าที่ยอมรับได้ (Accepted level) จึงจะหยุดฝึกหัด สุดท้ายค่าความถ่วงน้ำหนักที่ได้จะเป็นเหมือนฟังก์ชันที่ใช้ในการแปลงข้อมูล

2) การเรียนรู้แบบไม่มีผู้สอน (Unsupervised learning)

การเรียนรู้แบบไม่มีผู้สอนจะอาศัยชุดข้อมูลนำเข้าเพียงอย่างเดียวในการฝึกหัดโครงข่ายโดยไม่มีข้อมูลเป้าหมาย แต่จะใช้ข้อมูลนำออก (Output data) จากโครงข่ายแทน เมื่อป้อนข้อมูลเข้าโดยอาศัยค่าถ่วงน้ำหนักเป็นตัวแยกความแตกต่างของข้อมูลนำเข้าและนำไปเก็บไว้ในโหนดข้อมูลนำออกของโครงข่าย ซึ่งมีวัตถุประสงค์เพื่อใช้ในการจำแนกชุดข้อมูล (Clustering)

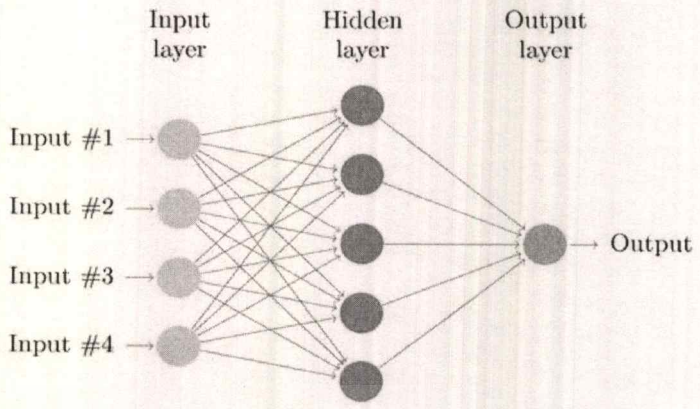
งานวิจัยชิ้นนี้ใช้การเรียนรู้แบบมีผู้สอน เนื่องจากโครงข่ายสามารถระบุกลุ่มของข้อมูลเป้าหมายได้อย่างแน่นอน ทำให้สะดวกในการออกแบบมากกว่าการเรียนรู้แบบไม่มีผู้สอน

2.5.3 การเชื่อมโยงของโครงข่ายประสาทเทียม (Neural network linking)

เพื่อให้โครงข่ายประสาทเทียมสามารถเรียนรู้ได้อย่างมีประสิทธิภาพ จำเป็นต้องมีการเชื่อมโยงของโครงข่ายได้ 2 ลักษณะคือ

1) โครงข่ายแบบส่งสัญญาณไปข้างหน้า (Feed-forward network)

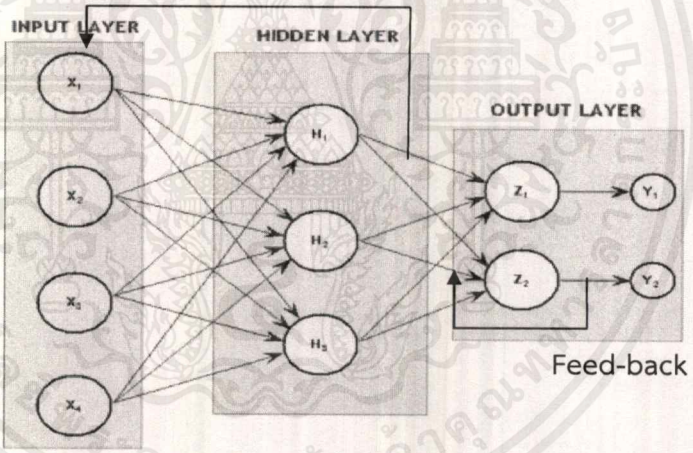
เป็นโครงข่ายที่การประมวลผลจะอาศัยชุดข้อมูลปัจจุบันและส่งค่าที่ประมวลผลได้ไปยังชั้นถัดๆ ไป กล่าวคือ โครงข่ายชนิดนี้จะประกอบด้วยชั้นต่างๆ โดยชั้นแรกจะเป็นชั้นนำเข้า (Input layer) และชั้นสุดท้ายเป็นชั้นนำออก (Output layer) ส่วนระหว่างชั้นนำเข้ากับชั้นนำออกอาจจะมีหรือไม่มีชั้นซ่อน (Hidden layers) อยู่ภายในก็ได้ ซึ่งขึ้นอยู่กับกฎการเรียนรู้ (Learning rule) ที่ใช้ในการสอนโครงข่าย เช่น ถ้าเป็นโครงข่ายเพอร์เซปตรอนแบบหลายชั้น (Multilayer Preceptron)จะมีชั้นซ่อนอยู่ระหว่างชั้นนำเข้ากับชั้นนำออกซึ่งอาจมีมากกว่าหนึ่งชั้นได้การเชื่อมต่อระหว่างชั้นของโครงข่ายแบบส่งสัญญาณไปข้างหน้าจะมีค่าถ่วงน้ำหนัก (weight) เป็นตัวเชื่อมและสัญญาณนำเข้าที่เข้ามาจะถูกส่งไปตามทิศทางของลูกศรจนถึงชั้นนำออกโดยไม่มีการป้อนกลับสามารถแสดงตัวแบบได้ดังรูปที่ 2-4



รูปที่ 2-4 ลักษณะโครงข่ายประสาทเทียมแบบส่งสัญญาณไปข้างหน้า

2) โครงข่ายแบบมีการป้อนกลับ (Feed-back network)

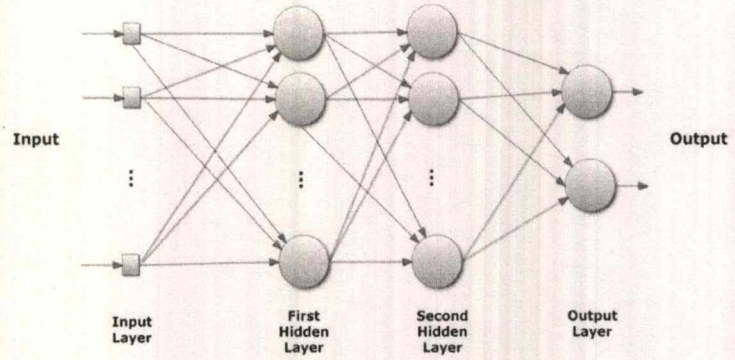
โครงข่ายชนิดนี้มีชื่อเรียกอีกอย่างหนึ่งว่า โครงข่ายหันกลับ (Recurrent network) เป็นโครงข่ายที่จะอาศัยทั้งข้อมูลในปัจจุบันและข้อมูลที่มีการประวิงเวลา มาใช้ในการประมวลผลของโครงข่ายประสาทเทียม สามารถแสดงตัวแบบโครงข่ายที่มีการป้อนกลับได้ดังรูปที่ 2-5



รูปที่ 2-5 ลักษณะโครงข่ายประสาทเทียมแบบมีการป้อนกลับ

2.5.4 การแพร่แบบย้อนกลับ (Back-propagation)

การแพร่แบบย้อนกลับเป็นขั้นตอนที่ใช้สอนโครงข่ายประสาทเทียมแบบเพอร์เซปตรอนหลายชั้น ซึ่งแบบจำลองโครงข่ายประสาทเทียมมีการเชื่อมโยงกันเป็นโครงข่ายแบบเป็นชั้นๆ โครงข่ายชนิดนี้มีการเชื่อมโยงกัน 3 ชั้น ประกอบด้วยชั้นนำเข้า (Input layer) ถัดมาเป็นชั้นซ่อน (Hidden layer) และชั้นสุดท้ายเป็นชั้นนำออก (Output layer) รูปที่ 2-6 แสดงโครงข่ายประสาทเทียมแบบเพอร์เซปตรอนหลายชั้นที่มีชั้นซ่อน 2 ชั้น



รูปที่ 2-6 โครงข่ายประสาทเทียมแบบเพอร์เซปตรอนหลายชั้น

ที่มาของชื่อการแพร่แบบย้อนกลับนั้นมาจากจุดที่ว่า วิธีการปรับค่าถ่วงน้ำหนักเพื่อให้ได้ค่าที่เหมาะสมนั้นจะใช้วิธีการสอนว่าค่าเป้าหมาย (Target) ของแต่ละข้อมูลนำเข้านั้นคืออะไร และใช้ค่าความผิดพลาด (Error) ของข้อมูลนำออกมาใช้เป็นตัวชี้นำในการปรับค่าถ่วงน้ำหนัก ดังนั้นการแพร่แบบย้อนกลับจึงเป็นกระบวนการเรียนรู้แบบมีผู้สอน แต่ปัญหาที่เกิดขึ้นคือ ไม่มีค่าเป้าหมายของสัญญาณที่ออกมาจากแต่ละเซลล์ประสาทในชั้นซ่อน ดังนั้นจึงต้องอาศัยการแพร่ความผิดพลาดจากชั้นนำออกกลับมายังชั้นซ่อนนั่นเอง

2.5.5 ปัจจัยที่ส่งผลต่อการเรียนรู้การแพร่แบบย้อนกลับ

1) การกำหนดค่าเริ่มต้นของค่าถ่วงน้ำหนัก

ก่อนที่จะทำการสอนโครงข่ายประสาทแบบหลายชั้น จำเป็นต้องกำหนดค่าเริ่มต้นให้กับค่าถ่วงน้ำหนักที่เชื่อมโยงระหว่างชั้นทุกชั้น โดยค่านี้จะเป็นเลขจำนวนจริงที่มีค่าน้อยๆ ที่ได้มาจากการสุ่มค่าเริ่มต้น (Randomness)

2) การกำหนดเกณฑ์การหยุดฝึกหัด

เกณฑ์ในการหยุดฝึกหัดนั้นขึ้นกับผู้ที่ทำการออกแบบโครงข่ายประสาทเทียม ว่าต้องการที่จะให้โครงข่ายประสาทเทียมมีความแม่นยำเพียงใด โดยทั่วไปนิยมใช้ค่าดัชนีที่ชี้ถึงค่าความผิดพลาดของระบบได้ ในงานวิจัยนี้ใช้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error, MSE)

3) อัตราการเรียนรู้(Learning rate, η)

อัตราการเรียนรู้เป็นค่าสัมประสิทธิ์ที่แสดงถึงการเรียนรู้ของโครงข่ายโดยทั่วไปค่าที่เหมาะสมจะอยู่ในช่วง 0.1 ถึง 0.5 ถ้าอัตราการเรียนรู้มีค่าสูง แสดงว่า กำหนดให้โครงข่ายมีการเปลี่ยนแปลงค่าถ่วงน้ำหนักที่มาก ในทางตรงกันข้ามถ้าอัตราการเรียนรู้มีค่าต่ำ แสดงว่ากำหนดให้โครงข่ายมีการเปลี่ยนแปลงค่าถ่วงน้ำหนักที่น้อย ซึ่งจำเป็นต้องใช้เวลาในการเรียนรู้ที่มากขึ้น แต่จะมีข้อดีคือโครงข่ายจะมีเสถียรภาพและไม่เกิดการแกว่ง (Oscillation) ขณะที่ทำการเรียนรู้

4) ค่าคงที่โมเมนตัม (Momentum onstant, α)

ค่าคงที่โมเมนตัมเป็นค่าสัมประสิทธิ์ที่ช่วยหน่วงไม่ให้การเปลี่ยนแปลงค่าถ่วงน้ำหนักนั้นมีค่ามากเกินไป เป็นการเพิ่มเสถียรภาพให้กับโครงข่ายได้อีกทางหนึ่ง ค่าโมเมนตัมที่เหมาะสมจะมีค่าเข้าใกล้ 1.0 และควรจะกำหนดให้สอดคล้องกับอัตราการเรียนรู้ด้วย เช่น ถ้ากำหนดอัตราการเรียนรู้ต่ำ ก็ควรที่จะมีค่าโมเมนตัมที่สูง ทำให้การเปลี่ยนแปลงค่าถ่วงน้ำหนักนั้นไม่มากจนเกินไป

การกำหนดโครงสร้างของโครงข่ายประสาทเทียม

จำนวนโหนดในชั้นข้อมูลเข้า โดยทั่วไปจะเท่ากับจำนวนค่าของตัวแปรอิสระ ส่วนการกำหนดจำนวนโหนดชั้นซ่อนยังไม่มีกฎเกณฑ์ที่แน่นอน โดยทั่วไปจำนวนโหนดในชั้นซ่อนจะได้จากการทดลอง สำหรับการกำหนดจำนวนโหนดในชั้นซ่อนจะใช้ Baum-Haussle rule ซึ่งได้เสนอไว้ในปี ค.ศ. 1998 เพื่อใช้ในการกำหนดจำนวนโหนดในชั้นซ่อน โดยคำนวณได้จากสมการที่ 2-5

$$n_{hidden} \leq n_{h\,max} = \frac{n_{data\,set} \times n_{input}}{n_{input} + n_{output}} \qquad (2\text{-}5)$$

โดยที่ n_{hidden} คือ จำนวนโหนดในชั้นซ่อน

$n_{h\,max}$ คือ จำนวนโหนดที่มากที่สุดในชั้นซ่อน

n_{input} คือ จำนวนโหนดให้ชั้นข้อมูลเข้า

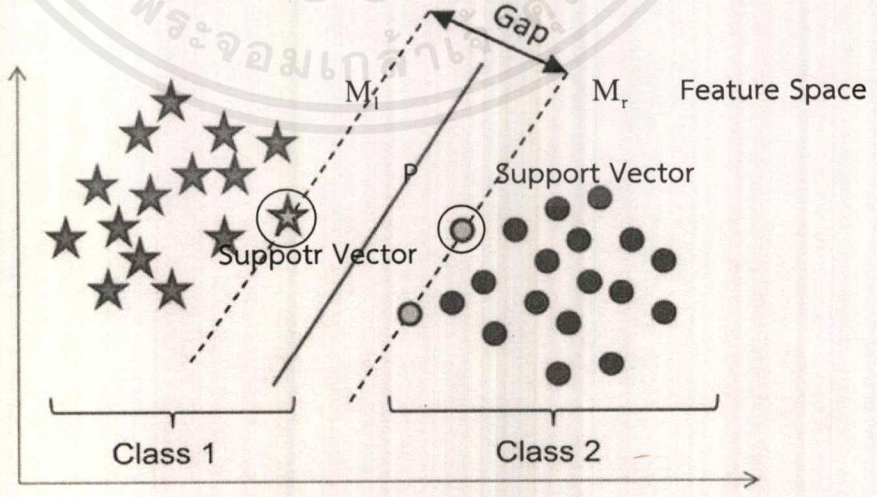
n_{output} คือ จำนวนโหนดในชั้นข้อมูลออก

$n_{data\,set}$ คือ จำนวนโหนดในการฝึกหัด

2.6 วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)

2.6.1 แนวความคิดของซัพพอร์ตเวกเตอร์แมชชีน (อานนท์ นามสนิท, 2549)

ซัพพอร์ตเวกเตอร์แมชชีนเป็นสมการที่ใช้ในการจำแนกค่าคุณลักษณะของ 2 กลุ่มที่วางตัวอยู่ในพื้นที่คุณลักษณะ (Feature Space) ออกจากกันโดยสร้างเส้นแบ่ง (Plane) ที่เป็นเส้นตรงขึ้นมา และเพื่อให้ทราบว่าเส้นตรงที่แบ่ง 2 กลุ่มออกจากกันนั้น เส้นตรงใดที่เป็นเส้นตรงที่ดีที่สุด โดยเส้นตรงนั้นจะเป็นเส้นขอบ (Margin) ออกไปทั้งสองข้าง โดยเส้นขอบที่เพิ่มนั้นจะขนานกับเส้นเดิมเสมอ เส้นขอบที่เพิ่มนั้นจะขยายออกไปจนกว่าจะสัมผัสกับค่าของกลุ่มตัวอย่างที่ใกล้ที่สุดดังรูปที่ 2-7



รูปที่ 2-7 การขยายตัวของเส้นขอบ

จากรูปที่ 2-7 เส้น M_l และ M_r คือเส้นขอบที่ขยายออกไปด้านซ้ายและขวาตามลำดับ และ P คือเส้นแบ่งข้อมูลทั้งสองกลุ่มเมื่อ M_l และ M_r ขยายออกไปจนไปสัมผัสค่าข้อมูลที่ใกล้ที่สุด ซึ่งข้อมูลที่อยู่บนเส้นขอบของทั้งสองฝั่งนั้นเรียกว่า ซัพพอร์ตเวกเตอร์ (Support vector) จะวัดค่าระยะความห่างของเส้นขอบโดยเส้น P จะเปลี่ยนความชันไปเรื่อยๆ เพื่อที่จะหาความกว้างสูงสุดของเส้นขอบ

กระบวนการโดยรวมของซัพพอร์ตเวกเตอร์แมชชีนนั้น เป็นการหาค่าความชันของเส้น P ที่มีขนาดของเส้นขอบสูงสุดนั่นเอง

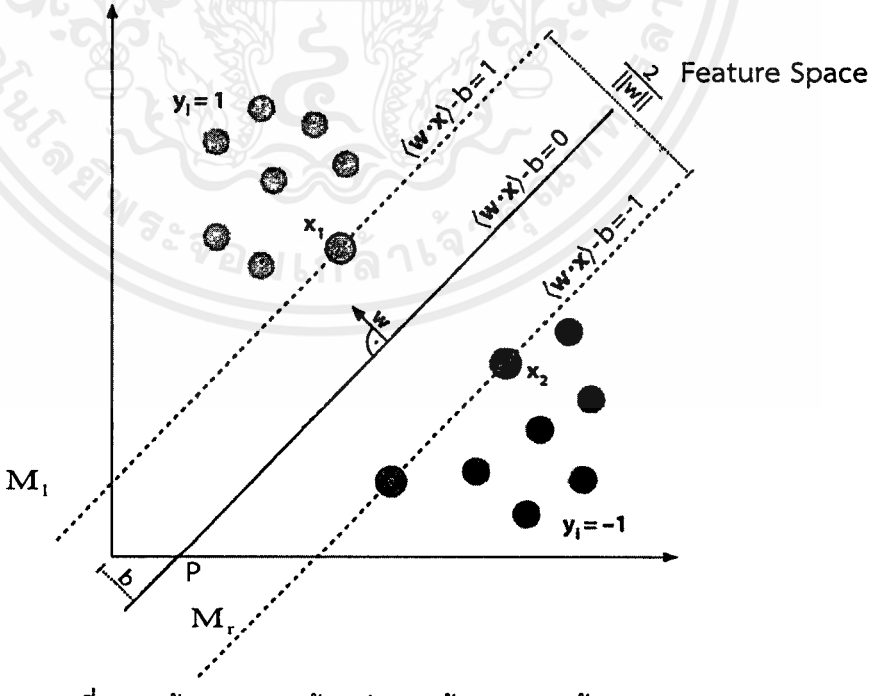
2.6.2 สมการพื้นฐานของซัพพอร์ตเวกเตอร์แมชชีน

ถ้านำแนวคิดของซัพพอร์ตเวกเตอร์แมชชีนที่กล่าวไปแล้วในข้อ 2.6.1 มาเขียนเป็นสมการเพื่อใช้ในการแก้ปัญหา โดยข้อมูลที่นำมาวางลงในพื้นที่คุณลักษณะนั้นเป็นกลุ่มข้อมูลที่อยู่ในรูปของเวกเตอร์

$$x = ((x_1, y_1),\ldots,(x_i, y_i)) \qquad (2\text{-}6)$$

เมื่อ x คือ ชุดค่าคุณลักษณะ

ค่าคุณลักษณะที่วางตัวอยู่ในพื้นที่คุณลักษณะ จะถูกแบ่งด้วยเส้นตรงดังรูปที่ 2-8 และเมื่อนำเส้นตรงมาแทนค่าด้วยสมการเส้นตรง $Y = mx + b$ โดยมีการกำหนดกลุ่มของข้อมูลทั้งสองฝั่งเป็นเพียงสองค่าที่ซึ่งแทนด้วยค่า y เพื่อให้ข้อมูลในที่อยู่กลุ่มเดียวกันที่มาจากหลายค่ากลายเป็นค่าเดียวดังสมการในรูปที่ 2-8



รูปที่ 2-8 เส้นขอบและเส้นแบ่งแทนด้วยสมการเส้นตรง

จากรูปที่ 2-8 เส้นตรง M_1 แทนด้วยสมการ $w^T x + b \geq y \geq 1$ ซึ่งข้อมูล y ที่มากกว่า 1 ก็จะถูกกำหนดค่าใหม่โดยให้ y เท่ากับ 1 และ พจน์ w ก็คือค่าความชัน เช่นเดียวกันกับ เส้นตรง M_r ที่ค่าของ y จะถูกกำหนดค่าใหม่เมื่อ y น้อยกว่า -1 เป็น -1 ดังนั้นสมการที่เกิดขึ้นใหม่จากสมการเส้นขอบ 2-7 และ 2-8 สามารถกำหนดได้ดังสมการที่ 2-9

เมื่อ $w^T x + b \geq y$ กำหนด $y = 1$ (2-7)

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

$y(w^T x + b) - 1 \geq 0$ (2-9)

โดย y คือค่ากลุ่มข้อมูล $(1,-1)$

w คือค่าความชัน

x คือค่าคุณลักษณะ

b คือค่าคงที่ (ค่าตัดแกน y)

2.6.3 ค่าความกว้างของเส้นขอบ (Margin)

การคำนวณความกว้างของเส้นขอบต้องทำการคำนวณพจน์ w ให้อยู่ในรูปปกติมาตรฐาน (Normalization) โดยคำนวณจากสมการที่ 2-7 ละ 2-8 เมื่อแทนค่า y ลงไป

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

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

$$w^T (x^+ - x^-) = 2$$

$$M = \left(\frac{w}{\|w\|} \right)^T (x^+ - x^-)$$

$$= \frac{2}{\|w\|} \quad (2\text{-}10)$$

$$W = \sum_{i=1}^{N} \alpha_i y_i x_i \quad (2\text{-}11)$$

โดยที่ M คือ ความกว้างของเส้นขอบ

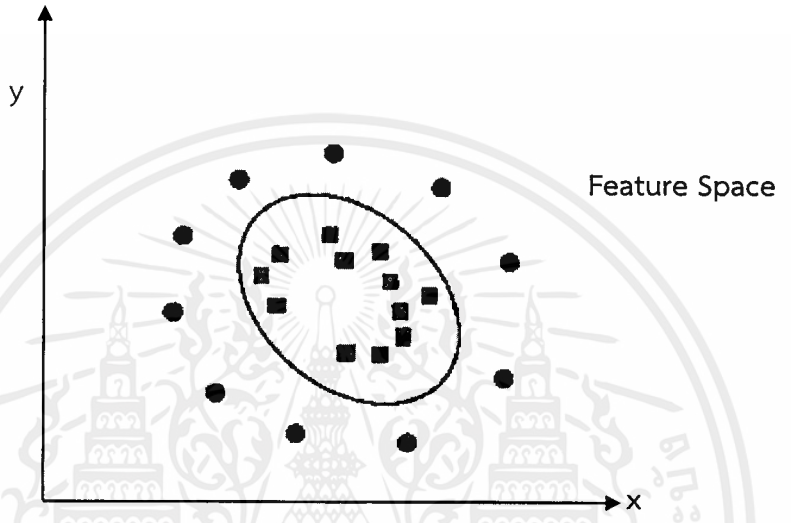
α คือ สัมประสิทธิ์คงที่

เมื่อนำค่า w ไปใส่ในสมการที่ 2-9 ซึ่งเป็นสมการในการหาเส้นแบ่ง จะได้

$$y_i \left(\sum_{i=1}^{N} \alpha_i y_i \left(x_i^T, x_j \right) + b \right) - 1 \geq 0 \quad (2\text{-}12)$$

2.6.4 เคอร์เนล (Kernel)

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



รูปที่ 2-9 รูปแบบการวางตัวที่ไม่สามารถแบ่งด้วยเส้นตรงได้

ในเคอร์เนลนั้นคือการคูณกันของชุดเวกเตอร์ของ x ใดๆ

$$K(x_i, x_j) = x_i^T x_j \qquad (2\text{-}13)$$

เคอร์เนลที่นิยมใช้มีอยู่ 2 ชนิดด้วยกัน คือ

1) โพลิโนเมียล (Polynomial)

$$K(x_i, x_j) = (< x_i, x_j > +1)^{d} \qquad (2\text{-}14)$$

เมื่อ d คือ ค่าเลขยกกำลัง

2) ฟังก์ชันเบสิสรัศมี (Radial Basis Function : RBF)

$$K(x_i, x_j) = exp\left[-\frac{x_i, x_i^2}{2\sigma^2}\right] \qquad (2\text{-}15)$$

เมื่อ σ คือ ค่าพารามิเตอร์

2.7 การเปรียบเทียบประสิทธิภาพของวิธีการจำแนกประเภท

การวัดประเมินผลมีความสำคัญเนื่องจากนำมาใช้ในการวิเคราะห์ประสิทธิภาพการทำงานของอัลกอริทึม (Kinzang, W., 2009) ตัวชี้วัดการประเมินผลสามารถที่จะพัฒนาจากเมทริกซ์ความสับสน (Confusion Matrix) แสดงให้เห็นในรูปที่ 2-10

		ผลการจำแนก	
		คำตอบเป็นบวก	คำตอบเป็นลบ
ค่าที่แท้จริง	คำตอบเป็นบวก	TP (True Positive)	FN (False Negative)
	คำตอบเป็นลบ	FP (False Positive)	TN (True Negative)

รูปที่ 2-10 เมทริกซ์ความสับสน (Confusion Matrix)

โดยที่ True Positive (TP) คือจำนวนข้อมูลที่จำแนกถูกว่าเป็นบวก

True Negative (TN) คือจำนวนข้อมูลที่จำแนกถูกว่าเป็นลบ

False Positive (FP) คือจำนวนข้อมูลที่จำแนกผิดว่าเป็นบวก ซึ่งค่าที่แท้จริงเป็นลบ

False Negative (FN) คือจำนวนข้อมูลที่จำแนกผิดว่าเป็นลบ ซึ่งค่าที่แท้จริงเป็นบวก

2.7.1 ค่าความถูกต้อง (Accuracy) คือ การแสดงผลการวัดที่ได้มีความถูกต้องในรูปอัตราส่วน

$$\text{Accuracy} = \frac{\text{จำนวนข้อมูลที่จำแนกถูกต้องค่าเป็นบวกและลบ}}{\text{จำนวนข้อมูลทั้งหมด}}$$

$$= \frac{TP + TN}{TP + TN + FP + FN} \qquad (2\text{-}16)$$

2.7.2 ค่าความแม่นยำ (Precision) คือ ความสามารถของเครื่องมือที่วัดได้แต่ละครั้งมีความแตกต่างของค่าวัดได้น้อยมาก เมื่อใช้เครื่องมือวัดนั้นไปวัดปริมาณตัวแปรเดิม

$$\text{Precision} = \frac{\text{จำนวนข้อมูลที่จำแนกถูกว่าเป็นบวก}}{\text{จำนวนข้อมูลที่ทำนายได้ว่าเป็นบวก}}$$

$$= \frac{TP}{TP + FP} \qquad (2\text{-}17)$$

2.7.3 ค่าความระลึก (Recall) คือ ค่าความระลึกของการจำแนกหมวดหมู่ของแต่ละกลุ่ม

$$\text{Recall} = \frac{\text{จำนวนข้อมูลที่จำแนกถูกว่าเป็นบวก}}{\text{จำนวนข้อมูลที่ค่าแท้จริงเป็นบวก}}$$

$$= \frac{TP}{TP + FN} \qquad (2\text{-}18)$$

2.7.4 ค่าความถ่วงดุล (F-Measure) ถ้าค่าความถ่วงดุลมีค่ามาก หมายถึงประสิทธิภาพในการจำแนกประเภทสูงด้วย

$$\text{F-Measure} = \frac{2 \times \text{ค่าความระลึก} \times \text{ค่าความแม่นยำ}}{\text{ค่าความระลึก} + \text{ค่าความแม่นยำ}}$$

$$= \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \qquad (2\text{-}19)$$

2.8 การเปรียบเทียบผลการทำนายของวิธีการจำแนกประเภท

ค่าการทำนาย (Prediction) คือส่วนที่แสดงผลการทำนายของแต่ละตัวอย่างของแต่ละชุดข้อมูล

ค่าการจำแนกได้ถูกต้อง (Correctly Classified Instances) คือ ค่าที่บอกว่าชุดข้อมูลมีอัตราการทำนายถูกต้องและผิดพลาดเท่าไร

เมทริกซ์ความสับสน (Confusion Matrix) เป็นรูปแบบตารางที่เฉพาะเจาะจงที่นำผลลัพธ์จากการทำนายมาใส่ในรูปตารางเมทริกซ์ ซึ่งจะช่วยให้ง่ายต่อการมองเห็นค่าทำนายของอัลกอริทึมดังรูปที่ 2-10

ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error : MSE) ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยใช้หลักการเดียวกันกับการหาค่าความแปรปรวนในทางสถิติ การวัดค่าความคลาดเคลื่อนด้วยวิธีนี้จะได้ค่าความคลาดเคลื่อนที่สูง เนื่องจากเป็นการนำความคลาดเคลื่อน ณ เวลาใดๆ มายกกำลังสองก่อนที่จะหาผลรวม แล้วจึงนำมาหาค่าเฉลี่ยอีกครั้งหนึ่ง นั่นคือ ค่า MSE ยิ่งน้อยหมายถึง การพยากรณ์ยิ่งแม่นยำ มีสูตรในการคำนวณดังนี้ [จุฑามาศ สิทธิโชคสถาพร, 2555]

$$MSE = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n} \qquad (2\text{-}20)$$

โดยที่ y_i แทน ค่าจริง

$\hat{y}_i$ แทน ค่าพยากรณ์

2.9 งานวิจัยที่เกี่ยวข้อง

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

เดช ธรรมศิริและ พยุง มีสัจ. (2554). กล่าวว่าในปัจจุบันการเพิ่มความแม่นยำสำหรับการจำแนกข้อมูลถือเป็นเรื่องสำคัญ ดังนั้น งานวิจัยครั้งนี้เป็นการนำเสนอวิธีการจำแนกข้อมูลด้วยวิธีร่วมกันตัดสินใจโดยใช้เทคนิคโครงข่ายประสาทเทียมร่วมกับการรวมกลุ่มตัดสินใจโดยใช้เทคนิคเอดาบูทเพื่อให้ความถูกต้องในการจำแนกข้อมูลที่สูงขึ้น ทดสอบบนฐานข้อมูลโรคเบาหวานได้แก่ Diabites Data จาก UCI การทดสอบความถูกต้องในการจำแนกข้อมูลDiabites Data พบว่า เทคนิคการร่วมกันตัดสินใจจากหลายโมเดลที่ผ่านการรวมกลุ่มตัดสินใจด้วยเทคนิคเอดาบูทมีผลลัพธ์ที่ดีกว่าการใช้เทคนิคแบบโมเดลเดี่ยว โดยผลการวิจัยพบว่าสำหรับข้อมูล Diabites Data ให้ผลลัพธ์ความถูกต้องสูงสุดที่ 75.02 % ในขณะที่โมเดลเดี่ยวให้ความแม่นยำ ที่น้อยกว่า

ภัทราวุฒิ แสงศิริ ศจีมาจ ณ วิเชียรและพยุงมีสัจ. (2552). กล่าวว่าการลดมิติของข้อมูลเป็นเทคนิคหนึ่งของการเตรียมข้อมูล (data preprocessing) สำหรับการทำเหมืองข้อมูล(data mining) สาเหตุเนื่องจากการทำเหมืองข้อมูลมีจำนวนมิติหรือตัวแปรมาก ทำให้ข้อมูลเกิดการกระจาย (data sparse) และทำให้เกิดปัญหามิติข้อมูล (Curse of Dimensionality) งานวิจัยนี้จะทำการเปรียบเทียบประสิทธิภาพความแม่นยำของเทคนิคการลดตัวแปรข้อมูลเข้า (input data) ระหว่างเทคนิคการเลือกตัวแปรแบบถอยหลังทีละขั้น (Backward Stepwise Feature Selection) และการวิเคราะห์องค์ประกอบ (Principle Component Analysis) โดยจะนำผลลัพธ์ที่ได้มาเป็นข้อมูลเข้าของโครงข่ายประสาทเทียม (Artificial Neural Network) เพื่อพยากรณ์กลุ่มข้อมูลโรคมะเร็งซึ่งผลการทดลองแสดงให้เห็นว่าการลดตัวแปรข้อมูลเข้าสาหรับโครงข่ายประสาทเทียมกับกลุ่มข้อมูลโรคมะเร็ง โดยใช้เทคนิคการเลือกตัวแปรแบบถอยหลังทีละขั้นมีความเหมาะสม คือให้ความแม่นยำ 94.62% และ 90.99% ต่างกับการวิเคราะห์องค์ประกอบที่ให้ค่าความแม่นยำ 98.32% และ 86.94%

นงเยาว์ ในอรุณ และ พรรณี สิทธิเดช. (2555). กล่าวถึงการจำแนกผู้ป่วยโรคหัวใจขาดเลือดและโรคหัวใจรูปแบบอื่นออกจากกันให้ชัดเจน เนื่องจากสองโรคนี้มีลักษณะอาการที่คล้ายกันทำให้แพทย์วินิจฉัยโรคได้ยาก ในปัจจุบันโรคหัวใจขาดเลือดเป็นโรคที่อันตรายและทำให้ผู้ป่วยทั่วโลกเสียชีวิตเป็นจำนวนมากและมีแนวโน้มเพิ่มขึ้นอย่างต่อเนื่อง ถ้าแพทย์สามารถตรวจพบโรคได้ก่อนจะช่วยลดความเสี่ยงของการเสียชีวิตของผู้ป่วยได้ วิธีการดำเนินการวิจัยได้รวบรวมข้อมูลผู้ป่วยจำนวน 2,500 ระเบียนมาทำการให้เป็นมาตรฐานเดียวกัน (Min-max normalization) และใช้อัลกอริทึม K-means เพื่อขจัดข้อมูลที่ไม่ชัดเจนออกไป ซึ่งได้ข้อมูลที่ถูกต้องสำหรับนำมาใช้ทดลองจำนวน 1,866 ระเบียน สุ่มแบ่งข้อมูลออกเป็นสองกลุ่ม กลุ่มละ 933 ระเบียน เพื่อนำมาสร้างโมเดลและใช้ทดสอบโมเดล ใช้ข้อมูลชุดแรกเพื่อสร้างโมเดลด้วยเทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support vector machine: SVM) และเทคนิคโครงข่ายประสาทเทียมแบบแพร่กลับ (Back propagation neural network: BPNN) เมื่อเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลทั้งสองวิธี พบว่าเทคนิค

SVM ได้ค่าความถูกต้อง (96.46%) ซึ่งมากกว่าเทคนิค BPNN (88.21%) สรุปได้ว่าเทคนิค SVM เป็นเทคนิคที่เหมาะสมในการจำแนกผู้ป่วยโรคหัวใจขาดเลือดออกจากโรคหัวใจรูปแบบอื่น

สมภพ ปฐมนพและคณะ. (2555). กล่าวว่า โรคเบาหวานเป็นโรคเรื้อรังที่ทำให้ผู้ป่วยมีคุณภาพชีวิตที่ลดลงเนื่องจากโรคเบาหวานมักจะก่อให้เกิดภาวะโรคแทรกซ้อนอื่นๆ ตามมาเช่น โรคหัวใจโรคความดันโลหิตสูง โรคระบบประสาทหรือแม้แต่การสูญเสียอวัยวะบางส่วนในร่างกายซึ่งเป็นสาเหตุร่วมของการเสียชีวิตด้วยโรคเบาหวาน งานวิจัยนี้ได้นำเสนอรูปแบบข้อมูลเชิงเวลาด้วยการเพิ่มคุณลักษณะข้อมูลเชิงเวลาจากข้อมูลประวัติการตรวจสุขภาพเพื่อการจากประเภทข้อมูลโดยการใช้อัลกอริทึม ได้แก่ Naïve Bayes, Logistic Regression, C4.5 (J48), Bagging และ SVMs ซึ่งงานวิจัยนี้ได้ทำการทดลองบนข้อมูลการตรวจสุขภาพในระหว่างปีพ.ศ. 2547–2553 (7 ปี) ของลูกจ้างโรงงงานอุตสาหกรรมในประเทศไทยโดยมีจำนวนลูกจ้างทั้งหมด 43,523 รายเป็นการตรวจเพียงครั้งเดียว 28,808 ราย และตรวจมากกว่าหนึ่งครั้ง 14,715 ราย โดยได้มีการทำรีแซมปลิงแบบแทนที่เพื่อปรับอินสแตนซ์ข้อมูลที่ใช้ในการเรียนในแต่ละคลาสให้สมดุลกันก่อนเข้าสู่กระบวนการเรียนต่อไป จากกลุ่มรายการตรวจสุขภาพ 3 รายการ คือ 1) การตรวจร่างกายทั่วไปโดยแพทย์ 2) การตรวจปัสสาวะ 3) การตรวจสารชีวเคมีในเลือด ผลการทดลองได้แสดงให้เห็นว่าข้อมูลที่เพิ่มคุณลักษณะข้อมูลเชิงเวลาให้ผลประสิทธิภาพการจำแนกประเภทดีกว่าข้อมูลแบบปกติที่ไม่มีคุณลักษณะข้อมูลเชิงเวลา

วาทินี นุ้ยเพียร และคณะ. (2553). ได้ทำการศึกษาเพื่อเปรียบเทียบวิธีการจำแนกข้อมูลโดยเลือกใช้อัลกอริทึมโครงข่ายประสาทเทียมแบบมัลติเลเยอร์เพอร์เซ็ปตรอน ซัพพอร์ตเวกเตอร์แมชชีน นาอีฟเบย์และเคเนียรเรสต์เนเบอร์ เพื่อประเมินประสิทธิภาพจากค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าความถ่วงดุล (F-Measure) ใช้ข้อมูลจาก UCI ประกอบด้วย Ozone Days และ Adult เลือกกลุ่มข้อมูล โดยมีจำนวนกลุ่ม (Class) เท่ากันในข้อมูลแต่ละชุด เป็นการทดลองแบบมีการเรียนรู้ จากผลการวิจัยอัลกอริทึมที่ดีที่สุดของข้อมูล Ozone Days คือซัพพอร์ตเวกเตอร์แมชชีน ฟังก์ชันเคอร์เนลแบบ Rbf มีค่าความถูกต้อง 95.83% ค่าความแม่นยำ 96% ค่าความระลึก 96% และค่าความถ่วงดุล 96% ส่วนข้อมูล Adult คือซัพพอร์ตเวกเตอร์แมชชีน ฟังก์ชันเคอร์เนลแบบ Polynomial มีค่าความถูกต้อง 79.66% ค่าความแม่นยำ 80% ค่าความระลึก 80% และค่าความถ่วงดุล 80%

วลัยลักษณ์ สุขสมบุญ และสมชาย ปราการเจริญ. (2553). ได้นำเสนอการเปรียบเทียบวิธีการจำแนกประเภทของปัญหาสำหรับ Helpdesk ด้วยซัพพอร์ตเวกเตอร์แมชชีน นาอีฟเบย์และเคเนียรเรสต์เนเบอร์ โดยนำข้อมูลมาตัดคำภาษาอังกฤษด้วยวิธีการอ้างอิงพจนานุกรมและคลังคำศัพท์เฉพาะที่ผู้วิจัยสร้างขึ้นใหม่ จากนั้นนำข้อมูลปัญหา Helpdesk แปลงให้อยู่ในรูปของเวกเตอร์ทางคณิตศาสตร์และทำการทดลองด้วยซัพพอร์ตเวกเตอร์แมชชีน นาอีฟเบย์และเคเนียรเรสต์เนเบอร์ โดยใช้การตรวจสอบไขว้กัน K-Fole = 10 โดยการทดลองปรับค่าพารามิเตอร์ของฟังก์ชันแกน Linear, Polynomial และ Radius Basic Function ในซัพพอร์ตเวกเตอร์แมชชีน นาอีฟเบย์และเคเนียร์เรสต์เนเบอร์ ผลการทดลองสามารถสรุปได้ว่าการทดลองเปรียบเทียบวิธีการจำแนกประเภทของปัญหาสำหรับ Helpdesk แสดงให้เห็นว่าเทคนิคที่นำเสนอด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนให้ค่าความถูกต้องที่ดีที่สุดโดยใช้ฟังก์ชันแกน Polynomial

พลอยพรรณ สอนสุวิทย์ และ ตรัสพงศ์ ไทยอุปถัมภ์. (2552). ได้นำเสนอการเปรียบเทียบประสิทธิภาพการตรวจจับสิ่งผิดปกติทางเครือข่าย โดยการสแกนหาจุดอ่อน ซึ่งเป็นการบุกรุกทางเครือข่ายที่มีความสำคัญ งานวิจัยนี้ได้ศึกษาเทคนิควิธีการจำแนก (Classification) มาตรวจจับการสแกน ได้แก่ SVM, C4.5, Naive Bayes และ Neural Network จากการศึกษาทำให้ทราบถึงประสิทธิภาพของแต่ละเทคนิค เช่น ร้อยละของความถูกต้อง (%Correct) และค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error) พบว่า วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ โดยใช้อัลกอริทึม C4.5 (J48) วิเคราะห์ได้แม่นยำที่สุดและมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด

บทที่ 3
วิธีการดำเนินงานวิจัย

3.1 การเก็บรวบรวมข้อมูล

ประชาการ คือ ผู้ป่วยที่เข้ามารับการรักษาที่โรงพยาบาลสีคิ้ว จังหวัด นครราชสีมา ตั้งแต่ปี 2499-2557 โดยตัวอย่างคือผู้ป่วยที่เข้ามารับการรักษาที่โรงพยาบาลสีคิ้ว จังหวัดนครราชสีมา ในช่วงเดือน มกราคม-มีนาคม พ.ศ. 2557 จำนวน 1,000 ระเบียน แบ่งคุณลักษณะของข้อมูลออกเป็น 13 คุณลักษณะ ดังรายละเอียดในภาคผนวก ก

1) เพศ (SEX)
2) กรรมพันธ์ (HEREDITY)
3) การสูบหรี่ (SMOKE)
4) การดื่มเครื่องดื่มแอลกอฮอล์ (ALCOHOL)
5) การออกกำลังกาย (EXERCISE)
6) การรับประทานอาหาร (EATING)
7) น้ำหนัก (WEIGH)
8) ส่วนสูง (HEIGHT)
9) ค่าดัชนีมวลกาย (BMI)
10) เกณฑ์ดัชนีมวลกาย (BMI MARK)
11) รอบเอว (WAISTLINE)
12) เกณฑ์ค่าความดันโลหิต (BLOOD PRESSURE)
13) การเป็นโรคความดันโลหิตสูง (HYPERTENSION)

3.2 เครื่องมือที่ใช้ในงานวิจัย

โปรแกรม WEKA (Waikato Environment for Knowledge Analysis) เวอร์ชั่น 3.7.5 ซึ่งเป็นโปรแกรมที่สามารถดาวน์โหลดได้จากเว็บไซต์ ซึ่งอยู่ภายใต้การควบคุมของ GPL License โปรแกรมนี้ถูกพัฒนาจากภาษาจาวาทั้งหมด ซึ่งเป็นที่นิยมในการใช้ทำเหมืองข้อมูล

3.3 วิธีการวิเคราะห์ข้อมูล

ในการทำงานวิจัยครั้งนี้ ผู้วิจัยได้เริ่มจากพิมพ์ข้อมูลการเป็นโรคความดันโลหิตสูงที่ได้จากโรงพยาบาลใส่ในโปรแกรม Microsoft Excel จำนวน 1,000 ระเบียน จากนั้นแบ่งข้อมูลออกเป็น 3 ส่วน โดยทำการสุ่มด้วยโปรแกรมสำเร็จรูป SPSS วิธีการสุ่มตัวอย่างแบบง่าย (Simple Random Sampling) โดยข้อมูลส่วนที่ 1 สุ่มข้อมูล 70 เปอร์เซ็นต์ของข้อมูลทั้งหมด จำนวน 700 ระเบียน ในการสร้างตัวแบบ ส่วนที่ 2 สุ่มข้อมูล 20 เปอร์เซ็นต์ของข้อมูลทั้งหมด จำนวน 200 ระเบียน ในการทดสอบความถูกต้องของตัวแบบ ส่วนที่ 3 สุ่มข้อมูล 10 เปอร์เซ็นของข้อมูลทั้งหมด จำนวน 100 ระเบียน ในการจำแนกและทำนายผล จากนั้นผู้วิจัยได้แปลงไฟล์ข้อมูลให้เป็นนามสกุล .CSV เพื่อใช้วิเคราะห์ประสิทธิภาพการจำแนกประเภทข้อมูลในโปรแกรม WEKA ซึ่งเป็นโปรแกรมที่สามารถนำมาทดสอบอัลกอรึทึมของวิธีการจำแนกประเภทได้ เนื่องจากมีอัลกอริทึมที่ระบุไว้ให้เลือกใช้ในโปรแกรมครบตามที่กำหนด ผู้วิจัยจึงได้กำหนดวิธีการจำแนกประเภทเพื่อนำมาทดสอบได้แก่

1. วิธีความใกล้เคียงกันมากที่สุด (K-Nearest Neighbor : KNN) โดยเลือกใช้อัลกอริทึม

1.1 IBk เป็นฟังก์ชั่นหลักที่น่าสนใจ ซึ่งเป็นพื้นฐานของอัลกอริทึม 8.1 อย่างไรก็ตาม อัลกอริทึม IBk ยังสามารถกำหนดน้ำหนักระยะห่างและทางเลือก (option) เพื่อกำหนดค่า k โดยใช้ Cross validation (Wu, X. and Kumar, V. 2009)

1.2 KStar เป็นตัวจำแนกกลุ่มโดยพิจารณาจากระเบียน (instance-based classifier) นั่นคือคำตอบ (class) ของตัวอย่างการทดสอบ (test instance) ขึ้นอยู่กับคำตอบของตัวอย่างฝึกหัด (training instance) หาได้โดยใช้ฟังก์ชันความคล้ายคลึง (similarity function) (สายชล สินสมบูรณ์ทอง, 2558)

1.3 LWL การใช้อัลกอริทึมโดยพิจารณาจากระเบียน (instance-based algorithm) เพื่อจัดน้ำหนักถ่วงให้แก่ระเบียนต่างๆ (สายชล สินสมบูรณ์ทอง, 2558)

2. วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ (Decision Tree) โดยเลือกใช้อัลกอริทึม

2.1 Decision Stump ใช้ในการวิเคราะห์การถดถอยและการจำแนกกลุ่ม ข้อมูลสูญหายพิจารณาในรูปค่าที่ใช้แบ่งแยก (separate value) (สายชล สินสมบูรณ์ทอง, 2558)

2.2 J48 หรือ C4.5 คืออัลกอริทึมที่ใช้ในการสร้างแผนภาพต้นไม้ เป็นที่นิยมโดยทั่วไป โดยเฉพาะผู้ศึกษาด้าน Machine Learning C4.5 พัฒนาโดย Quinlan, R. ในปี ค.ศ. 2009 (สายชล สินสมบูรณ์ทอง, 2558)

2.3 LMT ใช้สำหรับสร้างตัวแบบโลจิสติก ซึ่งเป็นต้นไม้จำแนกกลุ่มด้วยฟังก์ชันถดถอยโลจิสติกอยู่ที่ใบ อัลกอริทึมสามารถใช้กับตัวแปรเป้าหมายแบบไบนารีหรือแบบหลายคำตอบ สามารถใช้กับคุณลักษณะเชิงตัวเลข คุณลักษณะเชิงกลุ่มและข้อมูลสูญหาย (สายชล สินสมบูรณ์ทอง, 2558)

2.4 Random Forest คือ อัลกอริทึมที่ใช้หลายๆต้นไม้การตัดสินใจมาประมวลผลให้ความถูกต้องแม่นยำสูง จัดการข้อมูลได้มาก เหมาะสำหรับข้อมูลที่มีความสำคัญ (สายชล สินสมบูรณ์ทอง, 2558)

2.5 Random Tree คือ อัลกอริทึมที่ใช้ในการจำแนกหมวดหมู่เช่นเดียวกับ C4.5 โดยมีหลักการสร้างต้นไม้ในการสุ่มต้นไม้หลายๆแบบในแต่ละโหนดแล้วเลือกมาประมวลผลโดยไม่ใช้การตัด (Prune) (สายชล สินสมบูรณ์ทอง, 2558)

2.6 REP Tree คือ อัลกอริทึมที่มีหลักการสร้างต้นไม้จากค่าผลกำไรสารสนเทศ (Information Gain) การลดค่าความแปรปรวน (Variance Reduction) และการตัด คล้ายกับเทคนิคในอัลกอริทึม C4.5 แต่เพิ่มเทคนิคในการลดความผิดพลาดโดยการตัด เป็นอัลกอริทึมต้นไม้ที่มีจุดเด่นในด้านความเร็ว (สายชล สินสมบูรณ์ทอง, 2558)

3. วิธีโครงข่ายประสาทเทียม (Neural Networks) ใช้อัลกอริทึมเพอร์เซปตรอนหลายชั้น (Multilayer Perceptron) เป็นตัวจำแนกกลุ่ม (classifier) ซึ่งใช้การแพร่แบบย้อนกลับ (back propagation) เพื่อจำแนกระเบียน (instance) โครงข่ายนี้สามารถสร้างด้วยมือหรือสร้างด้วยอัลกอริทึมหรือทั้งสองอย่าง และสามารถควบคุมหรือดัดแปลงโครงข่าย ในระหว่างเวลาการฝึกหัดได้ โหนดในโครงข่ายนี้เป็น sigmoid ทั้งหมด ยกเว้นเมื่อคำตอบ (class) เป็นตัวเลข โหนดข้อมูลออกหรือผลลัพธ์ (output node) จะกลายเป็น unthresholded linear units (สายชล สินสมบูรณ์ทอง, 2558) ซึ่งในปัญหาพิเศษนี้จะกำหนดค่าอัตราการเรียนรู้ (Learning Rate) คือ 0.1, 0.2, 0.3, 0.4 และ 0.5 ค่าโมเมนตัม (Momentum) คือ 0.5, 0.6, 0.7, 0.8 และ 0.9 จำนวนรอบการสอน (TrainingTime) 20,000 รอบ การวิจัยครั้งนี้ใช้อัลกอริทึมของวิธีโครงข่ายประสาทเทียมแบบ Multilayer Perceptron โดยกำหนดให้มีชั้นซ่อน 1 ชั้น และหาจำนวนโหนดในชั้นซ่อนออกมาได้ 95 โหนด ดังนั้นจึงใช้ชั้นซ่อนที่มีจำนวนโหนดห่างกันเท่ากับ 10 ชั้นซ่อน คือ 10, 20, 30, 40, 50, 60, 70, 80, 90 และ 100 โหนด เพื่อให้ครอบคุมชั้นซ่อนที่หามาได้

การคำนวณค่าจำนวนโหนดในชั้นซ่อน ในที่นี้

$n_{dataset}$ คือ จำนวนโหนดในการฝึกหัด ในงานวิจัยนี้ใช้ข้อมูลในการฝึกหัด 100 โหนด

n_{output} คือ จำนวนโหนดในชั้นข้อมูลออก จะเท่ากับจำนวนผลลัพธ์มี 2 ผลลัพธ์ คือ ไม่เป็นโรค (Negative) และเป็นโรค (Positive)

n_{input} คือ จำนวนโหนดในชั้นข้อมูลเข้ามี 32 โหนด

ตัวแปร	คุณลักษณะ	จำนวนโหนดในชั้นข้อมูลเข้า
เพศ	1.ชาย 2.หญิง	2
กรรมพันธุ์	1.มี 2.ไม่มี 3.ไม่ทราบ	3
การสูบบุหรี่	1.สูบบุหรี่ 2.ไม่สูบบุหรี่ 3.เคยสูบแต่เลิกแล้ว	3
การดื่มแอลกอฮอล์	1.ดื่มแอลกอฮอล์ 2.ไม่ดื่มแอลกอออล์ 3.นานๆครั้ง 4.เคยดื่มแต่เลิกแล้ว	4
การออกกำลังกาย	1.ไม่ออกกำลังกาย 2.ออกกำลังกายน้อยกว่าสัปดาห์ละ 3 ครั้ง 3.ออกกำลังกายสัปดาห์ละ 3 ครั้ง 4.ออกกำลังกายมากกว่าสัปดาห์ละ 3 ครั้งๆละ 30 นาที 5.ออกกำลังกายทุกวันๆละ 30 นาที	5
การรับประทานอาหาร	1.หวาน 2.มัน 3.เค็ม 4.ไม่ชอบ	4
น้ำหนัก	1.เป็นตัวเลข	1
ส่วนสูง	1.เป็นตัวเลข	1
ค่าดัชนีมวลกาย	1.เป็นตัวเลข	1
เกณฑ์ดัชนีมวลกาย	1.น้ำหนักน้อยเกินไป 2.น้ำหนักปกติ 3.น้ำหนักเกิน 4.อ้วนระดับ 1 5.อ้วนระดับ 2	5

ตัวแปร	คุณลักษณะ	จำนวนโหนดในชั้นข้อมูลเข้า
รอบเอว	1.เป็นตัวเลข	1
เกณฑ์ค่าความดันโลหิต	1.ค่าความดันโลหิตอยู่ในเกณฑ์ปกติ 2.ค่าความดันโลหิตอยู่ในเกณฑ์มากกว่าปกติ	2
รวม		32

จำนวนโหนดมากที่สุดในชั้นซ่อน คือ

$$n_{hidden} \le n_{h\,max} = \frac{n_{data\,set} \times n_{input}}{n_{input} + n_{output}} = \frac{100 \times 32}{32 + 2}$$

$$= 94.12 \approx 95$$

4. วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) ใช้อัลกอริทึม SMO ใช้สำหรับฝึกหัดตัวจำแนกข้อมูลซัพพอร์ตเวกเตอร์ (Support Vector classifier) วิธีนี้ใช้การแทนที่ค่าข้อมูลสูญหายและแปลงข้อมูลคุณลักษณะเชิงกลุ่ม (nominal) ให้เป็นข้อมูลไบนารี่ (binary) นอกจากนี้ยังทำให้ข้อมูลคุณลักษณะทุกค่าอยู่ในรูปปกติมาตรฐาน (normalized) โดยอัลกอริทึม SMO แบ่งเป็นอัลกอริทึมย่อยๆ ได้แก่

4.1 Poly Kernel มีสมการเป็น $K(x, y) = (x, y)^p$ หรือ $K(x, y) = ((x, y)+1)^p$

4.2 Normalized Poly Kernel มีสมการเป็น $K(x, y) = \frac{(x, y)}{\sqrt{(x, x)(y, y)}}$

เมื่อ <x,y> = PolyKernel(x,y)

4.3 RBF Kernel มีสมการเป็น $K(x, y) = e^{-(\text{gamma} \times \langle x-y, x-y \rangle^2)}$

โดยที่ gamma คือ ค่า gamma = $\frac{1}{2\sigma^2}$ (จากสมการ 2-15)

4.4 Puk เคอร์เนลสากลโดยใช้ฟังก์ชันเพียรสัน VII (The Pearson VII function-based universal kernel.)

B. Uestuen, W.J. Melssen, L.M.C. Buydens (2006). ใช้การถดถอยซัพพอร์ตเวกเตอร์ (Support Vector Regression) พิจารณาเคอร์เนลสากลโดยใช้ฟังก์ชันเพียรสัน VII

การนำผลการวิเคราะห์มาประเมินผลเพื่อเปรียบเทียบประสิทธิภาพดังนี้

1. การเปรียบเทียบประสิทธิภาพในวิธีการจำแนกประเภท

นำข้อมูลที่ได้จากการวิเคราะห์มาประเมินผลเพื่อเปรียบเทียบว่าการจำแนกประเภทวิธีใดให้ค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุล การจำแนกประเภทข้อมูลสูงสุดระหว่างวิธีความใกล้เคียงกันมากที่สุด วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ วิธีโครงข่ายประสาทเทียม และวิธีซัพพอร์ทเวกเตอร์แมชชีน โดยค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุล มีค่ามากจะทำให้ประสิทธิภาพวิธีการจำแนกประเภทนั้นมีประสิทธิภาพสูง

2. การเปรียบเทียบผลการทำนาย

นำข้อมูลที่ได้จากการวิเคราะห์มาประเมินผลเพื่อเปรียบเทียบว่าการจำแนกประเภทวิธีใดให้ค่าความถูกต้องในการทำนายสูงสุด ระหว่างวิธีความใกล้เคียงกันมากที่สุด วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ วิธีโครงข่ายประสาทเทียม และวิธีซัพพอร์ทเวกเตอร์แมชชีน ซึ่งเปรียบเทียบผลของการทำนายจากค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error : MSE) โดยวิธีการจำแนกประเภทที่ให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยมีค่าต่ำ แสดงให้เห็นว่าวิธีการทำนายผลนั้นมีประสิทธิภาพสูง

บทที่ 4
ผลการวิเคราะห์ข้อมูล

งานวิจัยครั้งนี้ผู้วิจัยใช้การจำแนกประเภท โดยใช้เทคนิคการทำเหมืองข้อมูล โดยนำข้อมูลมาทำการสร้างตัวแบบ ทดสอบตัวแบบ และทำนายตัวแบบ โดยวิธีความใกล้เคียงกันมากที่สุด (K-Nearest Neighbor : KNN) โดยใช้อัลกอริทึม IBk, KStar และ LWL วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ (Decision Trees) โดยใช้อัลกอริทึม Decision Stump, J48, LMT, Random Forest, Random Tree และ REP Tree วิธีโครงข่ายประสาทเทียม (Neural Networks) โดยใช้อัลกอริทึม Multilayer Preceptron และวิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) โดยใช้อัลกอริทึม Normalized Poly Kernel, Poly Kernel, RBF Kernel และ Puk โดยนำมาเปรียบเทียบประสิทธิภาพในการจำแนกประเภท โดยพิจารณาค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) ค่าความถ่วงดุล (F-Measure) และเปรียบเทียบผลของการทำนาย โดยพิจารณาจากค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error : MSE) สรุปแยกตามวิธีการจำแนกประเภทดังนี้

4.1 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงโดยวิธีความใกล้เคียงกันมากที่สุด (K-Nearest Neighbor : KNN)

4.2 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงโดยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ (Decision Trees)

4.3 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงโดยวิธีโครงข่ายประสาทเทียม (Neural Networks)

4.4 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงโดยวิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)

4.5 สรุปผลวิธีการจำแนกประเภทที่มีประสิทธิภาพในการจำแนกประเภทและผลการทำนายที่ดีที่สุด

4.1 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงโดยวิธีความใกล้เคียงกันมากที่สุด (K-Nearest Neighbor : KNN)

4.1.1 ผลการวิเคราะห์ข้อมูลโดยอัลกอริทึม IBk

ตารางที่ 4-1 เมทริกซ์ความสับสนของข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยวิธีความใกล้เคียงกันมากที่สุดโดยอัลกอริทึม IBk

		ผลการจำแนก	
		ไม่เป็นโรค (Negative)	เป็นโรค (Positive)
ค่าที่แท้จริง	ไม่เป็นโรค (Negative)	72	13
	เป็นโรค (Positive)	8	7

จากตารางที่ 4-1 พบว่ามีข้อมูล 100 คน ตัวแบบสามารถจำแนกข้อมูลได้ถูกต้อง 79 คน โดยมีจำนวนข้อมูลที่จำแนกถูกว่าไม่เป็นโรค (Negative) มี 72 คน และจำนวนข้อมูลที่ถูกจำแนกถูกว่าเป็นโรค (Positive) มี 7 คน ตัวแบบทำนายข้อมูลไม่ถูกต้อง 21 คน โดยมีจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรค (Negative) ซึ่งค่าที่แท้จริงแล้วเป็นโรค (Positive) มี 8 คน และจำนวนข้อมูลที่จำแนกผิดว่าเป็นโรค (Positive) ซึ่งค่าที่แท้จริงแล้วไม่เป็นโรค (Negative) มี 13 คน โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ $(0.4576)^2 = 0.2094$ ซึ่งข้อมูลที่ทำนายผิดมีดังนี้ คือ คนที่ 1, 4, 5, 11, 15, 25, 27, 29, 32, 34, 39, 46, 52, 53, 62, 67, 68, 71, 75, 85, 87 ดังรายละเอียดในภาคผนวก ข (รูปที่ ข-3 ถึง ข-6)

4.1.2 ผลการวิเคราะห์ข้อมูลโดยอัลกอริทึม KStar

ตารางที่ 4-2 เมทริกซ์ความสับสนของข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยวิธีความใกล้เคียงกันมากที่สุดโดยอัลกอริทึม KStar

		ผลการจำแนก	
		ไม่เป็นโรค (Negative)	เป็นโรค (Positive)
ค่าที่แท้จริง	ไม่เป็นโรค (Negative)	77	8
	เป็นโรค (Positive)	11	4

จากตารางที่ 4-2 พบว่ามีข้อมูล 100 คน ตัวแบบสามารถทำนายข้อมูลได้ถูกต้อง 81 คน โดยมีจำนวนข้อมูลที่จำแนกถูกว่าไม่เป็นโรค (Negative) มี 77 คน และจำนวนข้อมูลที่ถูกจำแนกถูกว่าเป็นโรค (Positive) มี 4 คน ตัวแบบทำนายข้อมูลไม่ถูกต้อง 19 คน โดยมีจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรค (Negative) ซึ่งค่าที่แท้จริงเป็นโรค (Positive) มี 11 คน และจำนวนข้อมูลที่จำแนกผิดว่าเป็นโรค (Positive) ซึ่งค่าที่แท้จริงแล้วไม่เป็นโรค (Negative) มี 8 คน โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ $(0.3514)^2 = 0.1235$ ซึ่งข้อมูลที่ทำนายผิดมีดังนี้ คือ คนที่ 1, 8, 11, 12, 18, 24, 27, 29, 39, 46, 47, 53, 57, 58, 62, 68, 69, 81, 87 ดังรายละเอียดในภาคผนวก ข (รูปที่ ข-9 ถึง ข-12)

4.1.3 ผลการวิเคราะห์ข้อมูลโดยอัลกอริทึม LWL

ตารางที่ 4-3 เมทริกซ์ความสับสนของข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยวิธีความใกล้เคียงกันมากที่สุดโดยอัลกอริทึม LWL

		ผลการจำแนก	
		ไม่เป็นโรค (Negative)	เป็นโรค (Positive)
ค่าที่แท้จริง	ไม่เป็นโรค (Negative)	81	4
	เป็นโรค (Positive)	4	11

จากตารางที่ 4-3 พบว่ามีข้อมูล 100 คน ตัวแบบสามารถทำนายข้อมูลได้ถูกต้อง 92 คน โดยมีจำนวนข้อมูลที่จำแนกถูกว่าไม่เป็นโรค (Negative) มี 81 คน และจำนวนข้อมูลที่ถูกจำแนกถูกว่าเป็นโรค (Positive) มี 11 คน ตัวแบบทำนายข้อมูลไม่ถูกต้อง 8 คน โดยมีจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรค (Negative) ซึ่งค่าที่แท้จริงเป็น Positive มี 4 คน และจำนวนข้อมูลที่จำแนกผิดว่าเป็นโรค (Positive) ซึ่งค่าที่แท้จริงแล้วไม่เป็นโรค (Negative) มี 4 คน โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ $(0.2746)^2 = 0.0754$ ซึ่งข้อมูลที่ทำนายผิดมีดังนี้ คือ คนที่ ดังรายละเอียดในภาคผนวก ข (รูปที่ ข-15 ถึง ข-18)

ตารางที่ 4-4 ประสิทธิภาพการจำแนกประเภทของข้อมูลการเป็นโรคความดันโลหิตสูง โดยวิธีความใกล้เคียงกันมากที่สุดของแต่ละอัลกอริทึม

	ค่าการจำแนก			
อัลกอริทึม	ค่าความถูกต้อง	ค่าความแม่นยำ	ค่าความระลึก	ค่าความถ่วงดุล
IBk	79.00%	90.00%	84.71%	87.27%
KStar	81.00%	87.50%	90.59%	89.02%
LWL	92.00%	95.29%	95.29%	95.29%

จากตารางที่ 4-4 เมื่อพิจารณาจากค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุลประกอบกัน พบว่า อัลกอริทึม LWL มีค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุลมากที่สุด โดยแสดงตัวอย่างการคำนวณค่าการจำแนกดังตัวอย่างที่ 1 ในภาคผนวก ค จึงสรุปว่าวิธีการจำแนกประเภทที่ดีที่สุดที่ใช้สำหรับจำแนกประเภทข้อมูลการเป็นโรคดันโลหิตสูงของวิธีความใกล้เคียงกันมากที่สุด คือ อัลกอริทึม LWL

ตารางที่ 4-5 การทำนายผลของข้อมูลการเป็นโรคความดันโลหิตสูง โดยวิธีความใกล้เคียงกันมากที่สุดของแต่ละอัลกอริทึม

อัลกอริทึม	**ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย**
IBk	0.2094
KStar	0.1235
LWL	0.0754

จากตารางที่ 4-5 วิธีการจำแนกประเภทที่ดีที่สุดที่ใช้สำหรับทำนายข้อมูลการเป็นโรคความดันโลหิตสูงของวิธีความใกล้เคียงกันมากที่สุด คือ อัลกอริทึม LWL เนื่องจากมีค่าความคลาดเคลื่อนกำลังเฉลี่ยต่ำที่สุด

4.2 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงโดยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ (Decision Tree)

4.2.1 ผลการวิเคราะห์ข้อมูลโดยอัลกอริทึม Decision Stump

ตารางที่ 4-6 เมทริกซ์ความสับสนของข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจโดยอัลกอริทึม Decision Stump

		ผลการจำแนก	
		ไม่เป็นโรค (Negative)	เป็นโรค (Positive)
ค่าที่แท้จริง	ไม่เป็นโรค (Negative)	79	6
	เป็นโรค (Positive)	4	11

จากตารางที่ 4-6 พบว่ามีข้อมูล 100 คน ตัวแบบสามารถทำนายข้อมูลได้ถูกต้อง 90 คน โดยมีจำนวนข้อมูลที่จำแนกถูกว่าไม่เป็นโรค (Negative) มี 79 คน และจำนวนข้อมูลที่ถูกจำแนกถูกว่าเป็นโรค (Positive) มี 11 คน ตัวแบบทำนายข้อมูลไม่ถูกต้อง 10 คน โดยมีจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรค (Negative) ซึ่งค่าที่แท้จริงเป็นโรค (Positive) มี 4 คน และจำนวนข้อมูลที่จำแนกผิดว่าเป็นโรค (Positive) ซึ่งค่าที่แท้จริงแล้วไม่เป็นโรค (Negative) มี 6 คน โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ $(0.2822)^2 = 0.0796$ ซึงข้อมูลที่ทำนายผิดมีดังนี้ คือ คนที่ 4, 11, 16, 29, 32, 38, 75, 81, 85, 88 ดังรายละเอียดในภาคผนวก ข (รูปที่ ข-21 ถึง ข-24)

4.2.2 ผลการวิเคราะห์ข้อมูลโดยอัลกอริทึม J48

ตารางที่ 4-7 เมทริกซ์ความสับสนของข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจโดยอัลกอริทึม J48

		ผลการจำแนก	
		ไม่เป็นโรค (Negative)	เป็นโรค (Positive)
ค่าที่แท้จริง	ไม่เป็นโรค (Negative)	84	1
	เป็นโรค (Positive)	9	6

จากตารางที่ 4-7 พบว่ามีข้อมูล 100 คน ตัวแบบสามารถทำนายข้อมูลได้ถูกต้อง 90 คน โดยมีจำนวนข้อมูลที่จำแนกว่าถูกว่าไม่เป็นโรค (Negative) มี 84 คน และจำนวนข้อมูลที่ถูกจำแนกถูกว่าเป็นโรค (Positive) มี 6 คน ตัวแบบทำนายข้อมูลไม่ถูกต้อง 10 คน โดยมีจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรค (Negative) ซึ่งค่าที่แท้จริงแล้วเป็นโรค (Positive) มี 9 คน และจำนวนข้อมูลที่จำแนกผิดว่าเป็นโรค (Positive) ซึ่งค่าที่แท้จริงแล้วไม่เป็นโรค (Negative) มี 1 คน โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ $(0.296)^2 = 0.0876$ ซึ่งข้อมูลที่ทำนายผิดมีดังนี้ คือ คนที่ 6, 11, 24, 27, 29, 34, 38, 62, 75, 81 ดังรายละเอียดในภาคผนวก ข (รูปที่ ข-27 ถึง ข-30)

4.2.3 ผลการวิเคราะห์ข้อมูลโดยอัลกอริทึม LMT

ตารางที่ 4-8 เมทริกซ์ความสับสนของข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจโดยอัลกอริทึม LMT

		ผลการจำแนก	
		ไม่เป็นโรค (Negative)	เป็นโรค (Positive)
ค่าที่แท้จริง	ไม่เป็นโรค (Negative)	84	1
	เป็นโรค (Positive)	7	8

จากตารางที่ 4-8 พบว่ามีข้อมูล 100 คน ตัวแบบสามารถทำนายข้อมูลได้ถูกต้อง 92 คน โดยมีจำนวนข้อมูลที่จำแนกถูกว่าไม่เป็นโรค (Negative) มี 84 คน และจำนวนข้อมูลที่ถูกจำแนกถูกว่าเป็นโรค (Positive) มี 8 คน ตัวแบบทำนายข้อมูลไม่ถูกต้อง 8 คน โดยมีจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรค (Negative) ซึ่งค่าที่แท้จริงแล้วเป็นโรค (Positive) มี 7 คน และจำนวนข้อมูลที่จำแนกผิดว่าเป็นโรค (Positive) ซึ่งค่าที่แท้จริงแล้วไม่เป็นโรค (Negative) มี 1 คน โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย เท่ากับ $(0.2607)^2 = 0.0680$ ซึ่งข้อมูลที่ทำนายผิดมีดังนี้ คือ คนที่ 6, 11, 27, 29, 34, 38, 75, 81 ดังรายละเอียดในภาคผนวก ข (รูปที่ ข-33 ถึง ข-36)

4.2.4 ผลการวิเคราะห์ข้อมูลโดยอัลกอริทึม Random Forest

ตารางที่ 4-9 เมทริกซ์ความสับสนของข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจโดยอัลกอริทึม Random Forest

		ผลการจำแนก	
		ไม่เป็นโรค (Negative)	เป็นโรค (Positive)
ค่าที่แท้จริง	ไม่เป็นโรค (Negative)	80	5
	เป็นโรค (Positive)	9	6

จากตารางที่ 4-9 พบว่ามีข้อมูล 100 คน ตัวแบบสามารถทำนายข้อมูลได้ถูกต้อง 86 คน โดยมีจำนวนข้อมูลที่จำแนกถูกว่าไม่เป็นโรค (Negative) มี 80 คน และจำนวนข้อมูลที่ถูกจำแนกถูกว่าเป็นโรค (Positive) มี 6 คน ตัวแบบทำนายข้อมูลไม่ถูกต้อง 14 คน โดยมีจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรค (Negative) ซึ่งค่าที่แท้จริงแล้วเป็นโรค (Positive) มี 9 คน และจำนวนข้อมูลที่จำแนกผิดว่าเป็นโรค (Positive) ซึ่งค่าที่แท้จริงแล้วไม่เป็นโรค (Negative) มี 5 คน โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ $(0.289)^2 = 0.0835$ ซึ่งข้อมูลที่ทำนายผิดมีดังนี้ คือ คนที่ 1, 10, 11, 15, 27, 29, 34, 38, 53, 58, 62, 68, 75, 81 ดังรายละเอียดในภาคผนวก ข (รูปที่ ข-39 ถึง ข-42)

4.2.5 ผลการวิเคราะห์ข้อมูลโดยอัลกอริทึม Random Tree

ตารางที่ 4-10 เมทริกซ์ความสับสนของข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจโดยอัลกอริทึม Random Tree

		ผลการจำแนก	
		ไม่เป็นโรค (Negative)	เป็นโรค (Positive)
ค่าที่แท้จริง	ไม่เป็นโรค (Negative)	74	11
	เป็นโรค (Positive)	8	7

จากตารางที่ 4-10 พบว่ามีข้อมูล 100 คน ตัวแบบสามารถทำนายข้อมูลได้ถูกต้อง 81 คน โดยมีจำนวนข้อมูลที่จำแนกถูกว่าไม่เป็นโรค (Negative) มี 74 คน และจำนวนข้อมูลที่ถูกจำแนกถูกว่าเป็นโรค (Positive) มี 7 คน ตัวแบบทำนายข้อมูลไม่ถูกต้อง 19 คน โดยมีจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรค (Negative) ซึ่งค่าที่แท้จริงแล้วเป็นโรค (Positive) มี 8 คน และจำนวนข้อมูลที่จำแนกผิดว่าเป็นโรค (Positive) ซึ่งค่าที่แท้จริงแล้วไม่เป็นโรค (Negative) มี 11 คน โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ $(0.4284)^2 = 0.1835$ ซึ่งข้อมูลที่ทำนายผิดมีดังนี้ คือ คนที่ 6, 11, 15, 16, 20, 21, 24, 26, 27, 29, 31, 32, 38, 39, 53, 66, 68, 81, 87 ดังรายละเอียดในภาคผนวก ข (รูปที่ ข-45 ถึง ข-48)

4.2.6 ผลการวิเคราะห์ข้อมูลโดยอัลกอริทึม REP Tree

ตารางที่ 4-11 เมทริกซ์ความสับสนของข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจโดยอัลกอริทึม REP Tree

		ผลการจำแนก	
		ไม่เป็นโรค (Negative)	เป็นโรค (Positive)
ค่าที่แท้จริง	ไม่เป็นโรค (Negative)	83	2
	เป็นโรค (Positive)	8	7

จากตารางที่ 4-11 พบว่ามีข้อมูล 100 คน ตัวแบบสามารถทำนายข้อมูลได้ถูกต้อง 90 คน โดยมีจำนวนข้อมูลที่จำแนกถูกว่าไม่เป็นโรค (Negative) มี 83 คน และจำนวนข้อมูลที่ถูกจำแนกถูกว่าเป็นโรค (Positive) มี 7 คน ตัวแบบทำนายข้อมูลไม่ถูกต้อง 10 คน โดยมีจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรค (Negative) ซึ่งค่าที่แท้จริงแล้วเป็นโรค (Positive) มี 8 คน และจำนวนข้อมูลที่จำแนกผิดว่าเป็นโรค (Positive) ซึ่งค่าที่แท้จริงแล้วไม่เป็นโรค (Negative) มี 2 คน โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ $(0.2812)^2 = 0.0791$ ซึ่งข้อมูลที่ทำนายผิดมีดังนี้ คือ คนที่ 6, 11, 16, 24, 27, 29, 34, 38, 75, 81 ดังรายละเอียดในภาคผนวก ข (รูปที่ ข-51 ถึง ข-54)

ตารางที่ 4-12 ประสิทธิภาพการจำแนกประเภทของข้อมูลการเป็นโรคความดันโลหิตสูง โดยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจของแต่ละอัลกอริทึม

	ค่าการจำแนก			
อัลกอริทึม	ค่าความถูกต้อง	ค่าความแม่นยำ	ค่าความระลึก	ค่าความถ่วงดุล
Decision Stump	90.00%	95.18%	92.94%	94.05%
J48	90.00%	90.32%	**98.82%**	94.38%
LMT	**92.00%**	**92.31%**	**98.82%**	**94.45%**
Random Forest	86.00%	89.89%	94.21%	91.95%
Random Tree	81.00%	90.24%	87.06%	88.62%
REP Tree	90.00%	91.21%	97.65%	94.32%

จากตารางที่ 4-12 เมื่อพิจารณาจากค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุลประกอบกัน พบว่า อัลกอริทึม LMT มีค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุลมากที่สุด โดยแสดงตัวอย่างการคำนวณค่าการจำแนกดังตัวอย่างที่ 1 ในภาคผนวก ค จึงสรุปว่าวิธีการจำแนกประเภทที่ดีที่สุดที่ใช้สำหรับจำแนกประเภทข้อมูลการเป็นโรคดันโลหิตสูงของวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ คือ อัลกอริทึม LMT

ตารางที่ 4-13 การทำนายผลของข้อมูลการเป็นโรคความดันโลหิตสูง โดยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจของแต่ละอัลกอริทึม

อัลกอริทึม	**ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย**
Decision Stump	0.0796
J48	0.0876
LMT	**0.0680**
Random Forest	0.0835
Random Tree	0.1835
REP Tree	0.0791

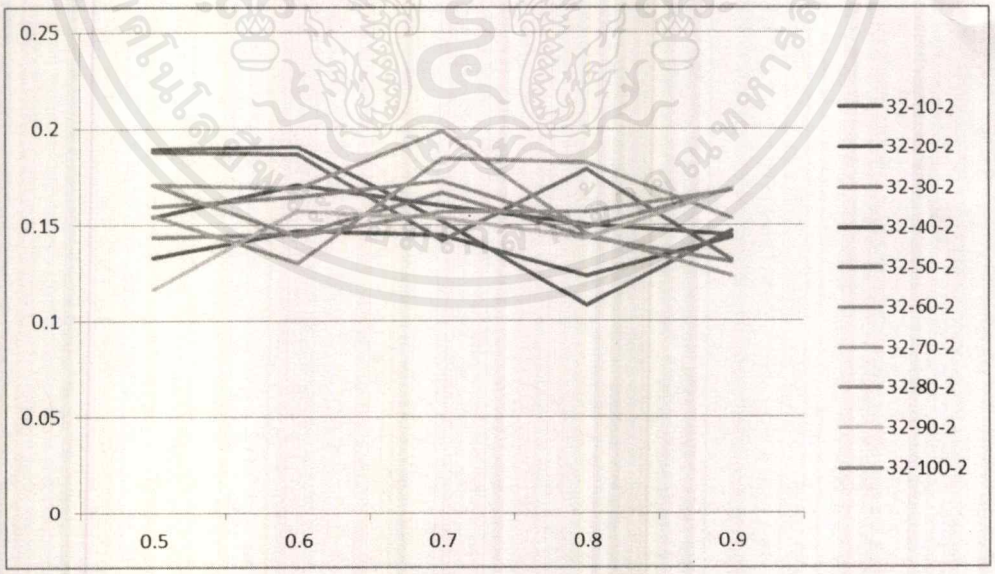
จากตารางที่ 4-5 วิธีการจำแนกประเภทที่ดีที่สุดที่ใช้สำหรับทำนายข้อมูลการเป็นโรคความดันโลหิตสูงของวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ คือ อัลกอริทึม LML เนื่องจากมีค่าความคลาดเคลื่อนกำลังเฉลี่ยต่ำที่สุด

4.3 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงโดยวิธีโครงข่ายประสาทเทียม (Neural Networks)

4.3.1 ผลการวิเคราะห์ข้อมูลโดยอัลกอริทึม Multilayer Preceptron

ตารางที่ 4-14 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย โดยกำหนดค่าอัตราการเรียนรู้ (Learning Rate) เป็น 0.1 จำนวนรอบการสอน (Training Time) 20,000 รอบ โดยวิธีโครงข่ายประสาทเทียม

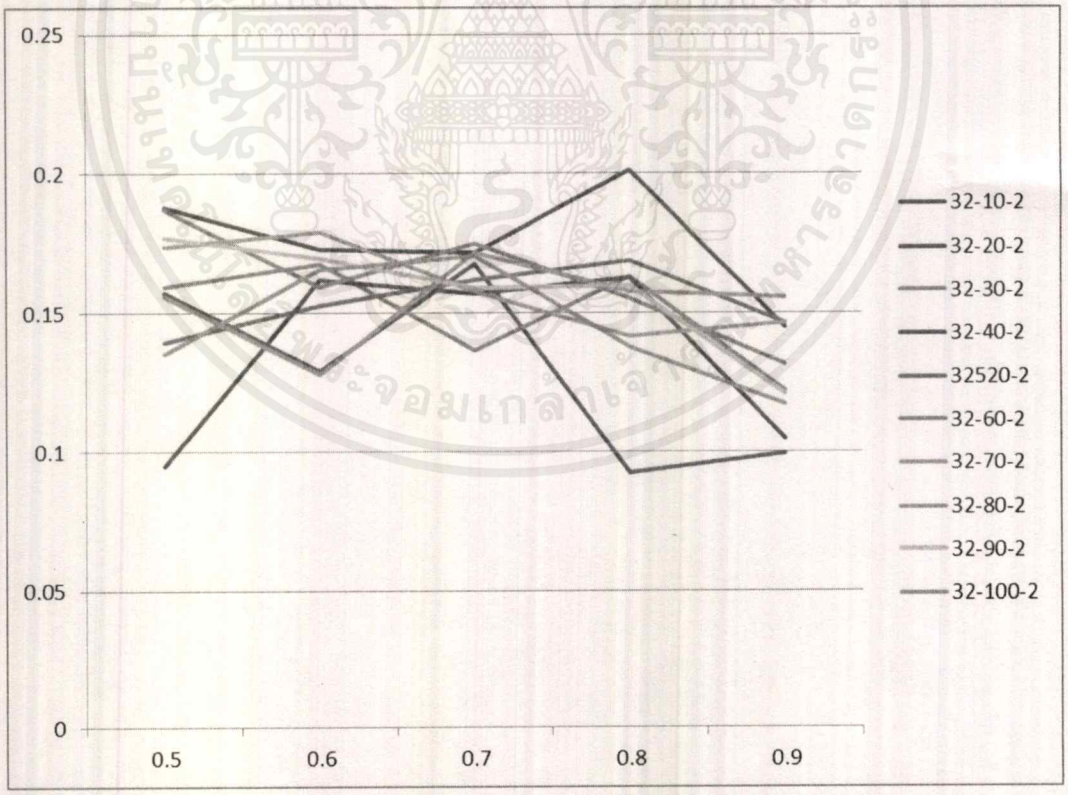
Input-Hidden-Output Node	ค่าโมเมนตัม (Momentum)				
	0.5	0.6	0.7	0.8	0.9
32-10-2	0.1898	0.1906	0.1519	0.1082	0.1471
32-20-2	0.1331	0.1467	0.1445	0.1238	0.1433
32-30-2	0.1438	0.1455	0.1669	0.1433	0.1308
32-40-2	0.1541	0.1709	0.1601	0.1504	0.1449
32-50-2	0.1880	0.1871	0.1417	0.1789	0.1318
32-60-2	0.1596	0.1646	0.1727	0.1488	0.1687
32-70-2	0.1548	0.1303	0.1845	0.1824	0.1538
32-80-2	0.1709	0.1689	0.1993	0.1453	0.1235
32-90-2	0.1166	0.1575	0.1534	0.1432	0.1690
32-100-2	0.1710	0.1444	0.1567	0.1567	0.1679



กราฟแสดงค่าความคลาดเคลื่อนกำลังสองเฉลี่ยที่อัตราการเรียนรู้เท่ากับ 0.1

ตารางที่ 4-15 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย โดยกำหนดค่าอัตราการเรียนรู้ (Learning Rate) เป็น 0.2 จำนวนรอบการสอน (Training Time) 20,000 รอบ โดยวิธีโครงข่ายประสาทเทียม

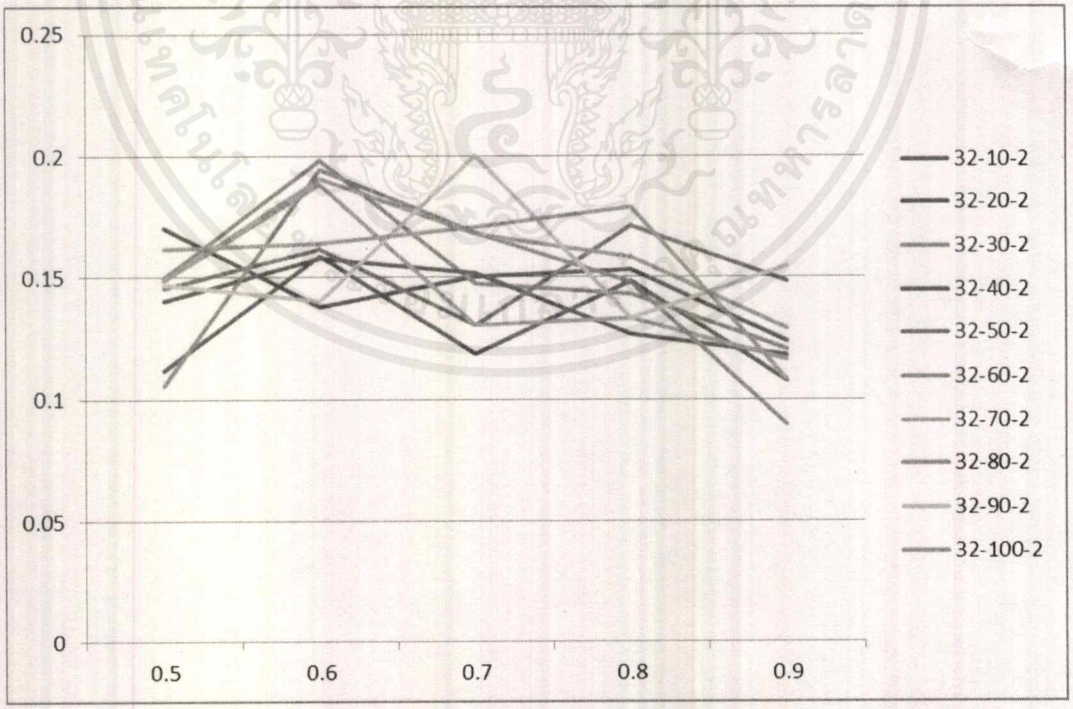
Input-Hidden-Output Node	ค่าโมเมนตัม(Momentum)				
	0.5	0.6	0.7	0.8	0.9
32-10-2	0.1575	0.1290	0.1676	0.0924	0.0995
32-20-2	0.1878	0.1731	0.1717	0.2012	0.1448
32-30-2	0.1592	0.1675	0.1368	0.1633	0.1221
32-40-2	0.0946	0.1619	0.1571	0.1627	0.1046
32-50-2	0.1395	0.1526	0.1617	0.1690	0.1466
32-60-2	0.1871	0.1595	0.1753	0.1551	0.1316
32-70-2	0.1740	0.1794	0.1577	0.1415	0.1465
32-80-2	0.1560	0.1279	0.1718	0.1578	0.1555
32-90-2	0.1771	0.1699	0.1589	0.1599	0.1210
32-100-2	0.1352	0.1657	0.1704	0.1377	0.1170



กราฟแสดงค่าความคลาดเคลื่อนกำลังสองเฉลี่ยที่อัตราการเรียนรู้เท่ากับ 0.2

ตารางที่ 4-16 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย โดยกำหนดค่าอัตราการเรียนรู้ (Learning Rate) เป็น 0.3 จำนวนรอบการสอน (Training Time) 20,000 รอบ โดยวิธีโครงข่ายประสาทเทียม

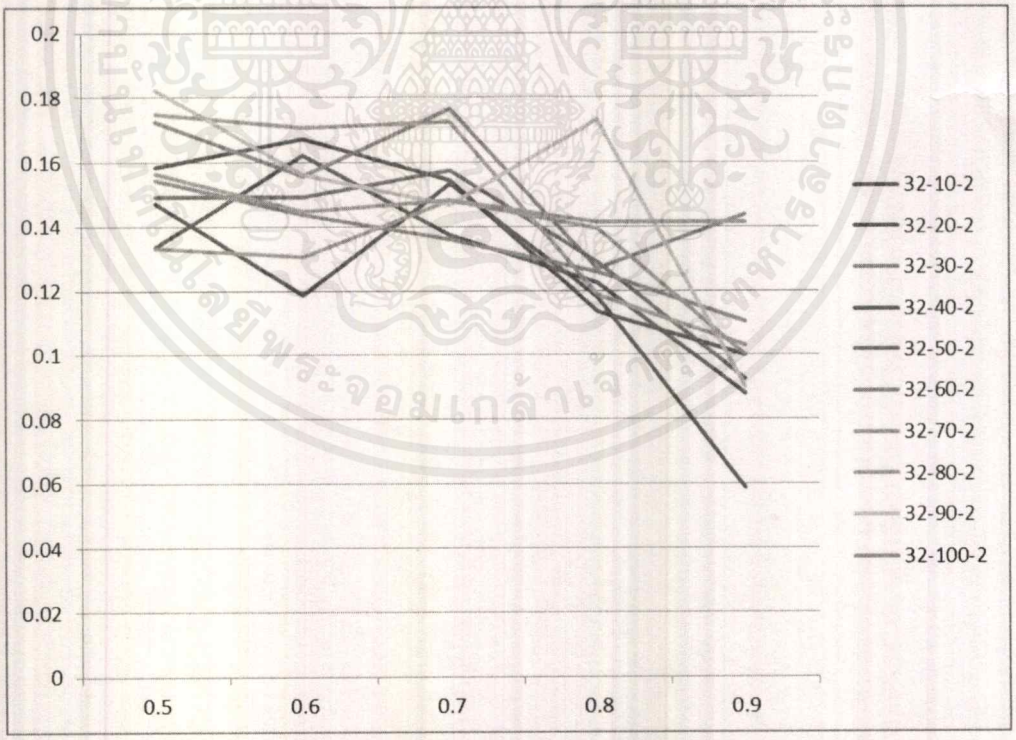
Input-Hidden-Output Node	ค่าโมเมนตัม(Momentum)				
	0.5	0.6	0.7	0.8	0.9
32-10-2	0.1120	0.1589	0.1188	0.1483	0.1075
32-20-2	0.1702	0.1379	0.1502	0.1533	0.1240
32-30-2	0.1500	0.1981	0.1474	0.1431	0.1207
32-40-2	0.1403	0.1581	0.1517	0.1265	0.1179
32-50-2	0.1455	0.1616	0.1307	0.1714	0.1484
32-60-2	0.1058	0.1945	0.1683	0.1481	0.0897
32-70-2	0.1492	0.1906	0.1681	0.1586	0.1290
32-80-2	0.1488	0.1882	0.1304	0.1332	0.1163
32-90-2	0.1470	0.1406	0.2001	0.1331	0.1548
32-100-2	0.1616	0.1640	0.1708	0.1788	0.1083



กราฟแสดงค่าความคลาดเคลื่อนกำลังสองเฉลี่ยที่อัตราการเรียนรู้เท่ากับ 0.3

ตารางที่ 4-17 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย โดยกำหนดค่าอัตราการเรียนรู้ (Learning Rate) เป็น 0.4 จำนวนรอบการสอน (Training Time) 20,000 รอบ โดยวิธีโครงข่ายประสาทเทียม

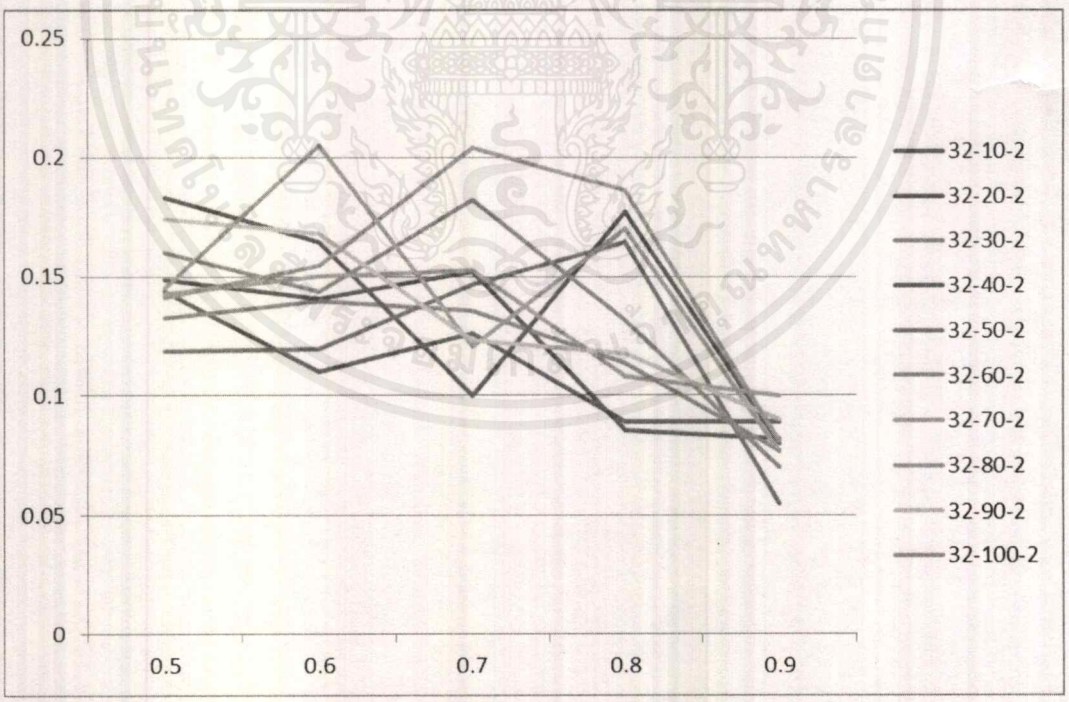
Input-Hidden-Output Node	ค่าโมเมนตัม(Momentum)				
	0.5	0.6	0.7	0.8	0.9
32-10-2	0.1336	0.1621	0.1373	0.1225	0.0880
32-20-2	0.1586	0.1674	0.1544	0.1134	0.1000
32-30-2	0.1726	0.1556	0.1767	0.1269	0.1440
32-40-2	0.1471	0.1188	0.1534	0.1178	0.0587
32-50-2	0.1492	0.1492	0.1576	0.1287	0.0925
32-60-2	0.1545	0.1439	0.1359	0.1257	0.1103
32-70-2	0.1750	0.1707	0.1724	0.1185	0.1029
32-80-2	0.1333	0.1305	0.1482	0.1393	0.1000
32-90-2	0.1824	0.1562	0.1470	0.1730	0.0900
32-100-2	0.1564	0.1449	0.1482	0.1414	0.1414



กราฟแสดงค่าความคลาดเคลื่อนกำลังสองเฉลี่ยที่อัตราการเรียนรู้เท่ากับ 0.4

ตารางที่ 4-18 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย โดยกำหนดค่าอัตราการเรียนรู้ (Learning Rate) เป็น 0.5 จำนวนรอบการสอน (Training Time) 20,000 รอบ โดยวิธีโครงข่ายประสาทเทียม

Input-Hidden-Output Node	ค่าโมเมนตัม(Momentum)				
	0.5	0.6	0.7	0.8	0.9
32-10-2	0.1433	0.1098	0.1258	0.0891	0.0891
32-20-2	0.1485	0.1407	0.1520	0.0854	0.0819
32-30-2	0.1325	0.1395	0.1354	0.1133	0.0768
32-40-2	0.1829	0.1643	0.1000	0.1768	0.0800
32-50-2	0.1185	0.1196	0.1465	0.1643	0.0548
32-60-2	0.1600	0.1435	0.1819	0.1316	0.0701
32-70-2	0.1425	0.1498	0.1523	0.1081	0.1000
32-80-2	0.1452	0.2049	0.1203	0.1700	0.0764
32-90-2	0.1739	0.1679	0.1228	0.1176	0.0899
32-100-2	0.1412	0.1544	0.2039	0.1861	0.0809



กราฟแสดงค่าความคลาดเคลื่อนกำลังสองเฉลี่ยที่อัตราการเรียนรู้เท่ากับ 0.5

ตารางที่ 4-19 เปรียบเทียบค่าความคลาดเคลื่อนกำลังสองเฉลี่ยที่น้อยที่สุดของแต่ละค่าอัตราการเรียนรู้

Input-Hidden-Output Node	ค่าอัตราการเรียนรู้	ค่าโมเมนตัม	ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย
32-10-2	0.1	0.8	0.1082
32-10-2	0.2	0.8	0.0924
32-60-2	0.3	0.9	0.0897
32-40-2	0.4	0.9	0.0587
32-50-2	0.5	0.9	**0.0548**

จากตารางที่ 4-19 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยที่มีค่าน้อยที่สุด คือ 0.0548 โดยมีค่าอัตราการเรียนรู้เท่ากับ 0.5 และค่าโมเมนตัมเท่ากับ 0.9 มีโหนดข้อมูลเข้า 32 โหนด โหนดซ่อน 50 โหนด และโหนดข้อมูลออก 2 โหนด ดังนั้น การทำนายตัวแบบจะกำหนดอัตราการเรียนรู้ 0.5 โมเมนตัม 0.9 โหนดซ่อน 50 โหนด และจำนวนรอบการสอน 20,000 รอบ ดังตารางที่ 4-20

ตารางที่ 4-20 เมทริกซ์ความสับสนของข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยวิธีโครงข่ายประสาทเทียม

		ผลการจำแนก	
		ไม่เป็นโรค (Negative)	**เป็นโรค** (Positive)
ค่าที่แท้จริง	**ไม่เป็นโรค** (Negative)	84	1
	เป็นโรค (Positive)	5	10

จากตารางที่ 4-20 พบว่ามีข้อมูล 100 คน ตัวแบบสามารถทำนายข้อมูลได้ถูกต้อง 94 คน โดยมีจำนวนข้อมูลที่จำแนกถูกว่าไม่เป็นโรค (Negative) มี 84 คน และจำนวนข้อมูลที่ถูกจำแนกถูกว่าเป็นโรค (Positive) มี 10 คน ตัวแบบทำนายข้อมูลไม่ถูกต้อง 6 คน โดยมีจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรค (Negative) ซึ่งค่าที่แท้จริงแล้วเป็นโรค (Positive) มี 5 คน และจำนวนข้อมูลที่จำแนกผิดว่าเป็นโรค (Positive) ซึ่งค่าที่แท้จริงแล้วไม่เป็นโรค (Negative) มี 1 คน โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ $(0.2341)^2 = 0.0548$ ซึ่งข้อมูลที่ทำนายผิดมีดังนี้ คือ คนที่ 11, 29, 38, 68, 75, 81 ดังรายละเอียดในภาคผนวก ข (รูปที่ ข-57 ถึง ข-60)

4.4 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงโดยวิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)

4.4.1 ผลการวิเคราะห์ข้อมูลโดยอัลกอริทึม Normalized Poly Kernel

ตารางที่ 4-21 เมทริกซ์ความสับสนของข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยอัลกอริทึม Normalized Poly Kernel

		ผลการจำแนก	
		ไม่เป็นโรค (Negative)	เป็นโรค (Positive)
ค่าที่แท้จริง	ไม่เป็นโรค (Negative)	84	1
	เป็นโรค (Positive)	7	8

จากตารางที่ 4-20 พบว่ามีข้อมูล 100 คน ตัวแบบสามารถทำนายข้อมูลได้ถูกต้อง 92 คน โดยมีจำนวนข้อมูลที่จำแนกถูกว่าไม่เป็นโรค (Negative) มี 84 คน และจำนวนข้อมูลที่ถูกจำแนกถูกว่าเป็นโรค (Positive) มี 8 คน ตัวแบบทำนายข้อมูลไม่ถูกต้อง 8 คน โดยมีจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรค (Negative) ซึ่งค่าที่แท้จริงแล้วเป็นโรค (Positive) มี 7 คน และจำนวนข้อมูลที่จำแนกผิดว่าเป็นโรค (Positive) ซึ่งค่าที่แท้จริงแล้วไม่เป็นโรค (Negative) มี 1 คน โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ $(0.2828)^2 = 0.0800$ ซึ่งข้อมูลที่ทำนายผิดมีดังนี้ คือ คนที่ 6, 11, 27, 29, 34, 38, 75, 81 ดังรายละเอียดในภาคผนวก ข (รูปที่ ข-63 ถึง ข-66)

4.4.2 ผลการวิเคราะห์ข้อมูลโดยอัลกอริทึม Poly Kernel

ตารางที่ 4-22 เมทริกซ์ความสับสนของข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยอัลกอริทึม Poly Kernel

		ผลการจำแนก	
		ไม่เป็นโรค (Negative)	เป็นโรค (Positive)
ค่าที่แท้จริง	ไม่เป็นโรค (Negative)	83	2
	เป็นโรค (Positive)	8	7

จากตารางที่ 4-21 พบว่ามีข้อมูล 100 คน ตัวแบบสามารถทำนายข้อมูลได้ถูกต้อง 90 คน โดยมีจำนวนข้อมูลที่จำแนกถูกว่าไม่เป็นโรค (Negative) มี 83 คน และจำนวนข้อมูลที่ถูกจำแนกถูกว่าเป็นโรค (Positive) มี 7 คน ตัวแบบทำนายข้อมูลไม่ถูกต้อง 10 คน โดยมีจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรค (Negative) ซึ่งค่าที่แท้จริงแล้วเป็นโรค (Positive) มี 8 คน และจำนวนข้อมูลที่จำแนกผิดว่าเป็นโรค (Positive) ซึ่งค่าที่แท้จริงแล้วไม่เป็นโรค (Negative) มี 2 คน โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ $(0.3162)^2 = 0.1000$ ซึ่งข้อมูลที่ทำนายผิดมีดังนี้ คือ คนที่ 6, 11, 16, 24, 27, 29, 34, 38, 75, 81 ดังรายละเอียดในภาคผนวก ข (รูปที่ ข-69 ถึง ข-72)

4.4.3 ผลการวิเคราะห์ข้อมูลโดยอัลกอริทึม RBF Kernel

ตารางที่ 4-23 เมทริกซ์ความสับสนของข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยอัลกอริทึม RBF Kernel

		ผลการจำแนก	
		ไม่เป็นโรค (Negative)	เป็นโรค (Positive)
ค่าที่แท้จริง	ไม่เป็นโรค (Negative)	85	0
	เป็นโรค (Positive)	15	0

จากตารางที่ 4-23 พบว่ามีข้อมูล 100 คน ตัวแบบสามารถทำนายข้อมูลได้ถูกต้อง 85 คน โดยมีจำนวนข้อมูลที่จำแนกถูกว่าเป็นโรค (Negative) มี 85 คน ตัวแบบทำนายข้อมูลไม่ถูกต้อง 15 คน โดยมีจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรค (Negative) ซึ่งค่าที่แท้จริงแล้วไม่เป็นโรค (Positive) มี 15 คน โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ $(0.3873)^2 = 0.1500$ ซึ่งข้อมูลที่ทำนายผิดมีดังนี้ คือ คนที่ 6, 7, 11, 12, 24, 27, 29, 34, 47, 53, 57, 62, 68, 75, 81 ดังรายละเอียดในภาคผนวก ข (รูปที่ ข-75 ถึง ข-78)

4.4.4 ผลการวิเคราะห์ข้อมูลโดยอัลกอริทึม Puk

ตารางที่ 4-24 เมทริกซ์ความสับสนของข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยอัลกอริทึม Puk

		ผลการจำแนก	
		ไม่เป็นโรค (Negative)	เป็นโรค (Positive)
ค่าที่แท้จริง	ไม่เป็นโรค (Negative)	82	3
	เป็นโรค (Positive)	14	1

จากตารางที่ 4-24 พบว่ามีข้อมูล 100 คน ตัวแบบสามารถทำนายข้อมูลได้ถูกต้อง 83 คน โดยมีจำนวนข้อมูลที่จำแนกถูกว่าไม่เป็นโรค (Negative) มี 82 คน และจำนวนข้อมูลที่ถูกจำแนกถูกว่าเป็นโรค (Positive) มี 1 คน ตัวแบบทำนายข้อมูลไม่ถูกต้อง 17 คน โดยมีจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรค (Negative) ซึ่งค่าที่แท้จริงแล้วเป็นโรค (Positive) มี 14 คน และจำนวนข้อมูลที่จำแนกผิดว่าเป็นโรค (Positive) ซึ่งค่าที่แท้จริงแล้วไม่เป็นโรค (Negative) มี 3 คน โดยมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ $(0.4123)^2 = 0.1700$ ซึ่งข้อมูลที่ทำนายผิดมีดังนี้ คือ คนที่ 6, 7, 11, 12, 15, 24, 27, 29, 34, 47, 53, 62, 67, 68, 71, 75, 81 ดังรายละเอียดในภาคผนวก ข (รูปที่ ข-81 ถึง ข-84)

ตารางที่ 4-25 ประสิทธิภาพการจำแนกประเภทของข้อมูลการเป็นโรคความดันโลหิตสูง โดยวิธีซัพพอร์ตเวกเตอร์แมชชีนของแต่ละอัลกอริทึม

	ค่าการจำแนก			
อัลกอริทึม	ค่าความถูกต้อง	ค่าความแม่นยำ	ค่าความระลึก	ค่าความถ่วงดุล
Normalized Poly Kernel	**92.00%**	**92.31%**	98.82%	**95.45%**
Poly Kernel	90.00%	91.21%	97.65%	94.32%
RBF Kernel	85.00%	85.00%	**100.00%**	91.89%
Puk	83.00%	85.42%	96.47%	90.61%

จากตารางที่ 4-25 เมื่อพิจารณาจากค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุลประกอบกัน พบว่า อัลกอริทึม Normalized Poly Kernel มีค่าการจำแนกที่สูงที่สุดมีจำนวนมากที่สุด ได้แก่ ค่าความถูกต้อง ค่าความแม่นยำ และค่าความถ่วงดุลมากที่สุด ส่วนค่าความระลึกถึงแม้จะไม่ได้มากที่สุด แต่ค่าก็แตกต่างจากค่าความระลึกที่มากที่สุดเพียงเล็กน้อย โดยแสดงตัวอย่างการคำนวณค่าการจำแนกดังตัวอย่างที่ 1 ในภาคผนวก ค จึงสรุปว่าวิธีการจำแนกประเภทที่ดีที่สุดที่ใช้สำหรับจำแนกประเภทข้อมูลการเป็นโรคดันโลหิตสูงของวิธีซัพพอร์ตเวกเตอร์แมชชีน คือ อัลกอริทึม Normalized Poly Kernel

ตารางที่ 4-26 การทำนายผลของข้อมูลการเป็นโรคความดันโลหิตสูง โดยวิธีซัพพอร์ตเวกเตอร์แมชชีนของแต่ละอัลกอริทึม

อัลกอริทึม	**ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย**
Normalized Poly Kernel	0.0800
Poly Kernel	0.1000
RBF Kernel	0.1500
Puk	0.1700

จากตารางที่ 4-26 วิธีการจำแนกประเภทที่ดีที่สุดที่ใช้สำหรับทำนายข้อมูลการเป็นโรคความดันโลหิตสูงของวิธีซัพพอร์ตเวกเตอร์แมชชีน คือ อัลกอริทึม Normalized Poly Kernel เนื่องจากมีค่าความคลาดเคลื่อนกำลังเฉลี่ยต่ำที่สุด

4.5 สรุปผลวิธีการจำแนกประเภทที่มีประสิทธิภาพการจำแนกประเภทและผลการทำนายที่ดีที่สุด

ตารางที่ 4-27 ประสิทธิภาพการจำแนกประเภทของการเป็นโรคความดันโลหิตสูง

อัลกอริทึม	ค่าการจำแนก			
	ค่าความถูกต้อง	ค่าความแม่นยำ	ค่าความระลึก	ค่าความถ่วงดุล
วิธีความใกล้เคียงกันมากที่สุด (LWL)	92.00%	**95.29%**	95.29%	95.29%
วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ (LMT)	90.00%	92.31%	**98.82%**	95.45%
วิธีโครงข่ายประสาทเทียม $(\eta = 0.5, \alpha = 0.9)$	**94.00%**	94.38%	**98.82%**	**96.55%**
วิธีซัพพอร์ตเวกเตอร์แมชชีน (Normalized Poly Kernel)	92.00%	92.31%	**98.82%**	95.45%

จากตารางที่ 4-27 เมื่อพิจารณาจากค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุลประกอบกัน พบว่า วิธีโครงข่ายประสาทเทียม มีค่าการจำแนกสูงที่สุดมีจำนวนมากที่สุด คือ ค่าความถูกต้อง ค่าความระลึก และค่าความถ่วงดุลมากที่สุด ส่วนค่าความแม่นยำถึงแม้จะไม่ได้สูงที่สุด แต่ค่าก็แตกต่างจากค่าความแม่นยำที่สูงที่สุดเพียงเล็กน้อย จึงสรุปว่าวิธีการจำแนกประเภทที่ดีที่สุดที่ใช้สำหรับจำแนกประเภทข้อมูลการเป็นโรคดันโลหิตสูง คือ วิธีโครงข่ายประสาทเทียม $(\eta = 0.5, \alpha = 0.9)$

ตารางที่ 4-28 การทำนายผลของข้อมูลการเป็นโรคความดันโลหิตสูง

อัลกอริทึม	ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย
วิธีความใกล้เคียงกันมากที่สุด (LWL)	0.0754
วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ (LMT)	0.0680
วิธีโครงข่ายประสาทเทียม $(\eta = 0.5, \alpha = 0.9)$	**0.0548**
วิธีซัพพอร์ตเวกเตอร์แมชชีน (Normalized Poly Kernel)	0.0800

จากตารางที่ 4-28 วิธีการจำแนกประเภทที่ดีที่สุดที่ใช้สำหรับทำนายข้อมูลการเป็นโรคความดันโลหิตสูง คือ วิธีโครงข่ายประสาทเทียม $(\eta = 0.5, \alpha = 0.9)$ เนื่องจากมีค่าความคลาดเคลื่อนกำลังเฉลี่ยต่ำที่สุด

บทที่ 5

สรุปผล อภิปรายผล และข้อเสนอแนะ

5.1 สรุปผลการวิจัย

ในการจัดทำวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาทำความเข้าใจและเปรียบเทียบประสิทธิภาพวิธีการจำแนกประเภท รวมทั้งเปรียบเทียบผลการทำนายของวิธีการจำแนกประเภทของการเป็นโรคความดันโลหิตสูง วิธีการจำแนกประเภทที่นำมาเปรียบเทียบประสิทธิภาพ คือ วิธีความใกล้เคียงกันมากที่สุด โดยใช้อัลกอริทึม IBK, KStar และ LWL ซึ่งอัลกอริทึมที่มีประสิทธิภาพดีที่สุดของวิธีความใกล้เคียงกันมากที่สุดในการทำวิจัยครั้งนี้ คือ อัลกอริทึม LWL วิธีแผนภาพต้นไม้เพื่อการตัดสินใจโดยใช้อัลกอริทึม J48, Decision Stump, LMT, Random Forest, Random Tree และ REP Tree ซึ่งอัลกอริทึมที่มีประสิทธิภาพดีที่สุดของวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ ในการทำวิจัยครั้งนี้ คือ อัลกอริทึม LMT วิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึมชนิด Multilayer Perceptron โดยกำหนดค่าอัตราการเรียนรู้เป็น 0.1, 0.2, 0.3, 0.4 และ 0.5 ค่าโมเมนตัม เป็น 0.5, 0.6, 0.7, 0.8 และ 0.9 จำนวนรอบการสอน 20,000 รอบ ใช้ชั้นซ่อนที่มีจำนวนโหนดห่างกันเท่ากับ 10 ชั้นซ่อน คือ 10, 20, 30, 40, 50, 60, 70, 80, 90 และ 100 โหนด ประสิทธิภาพในการจำแนกดีที่สุดของวิธีโครงข่ายประสาทเทียมโดยใช้ชนิด Multilayer Perceptron ในการทำวิจัยครั้งนี้ คือ อัตราการเรียนรู้ 0.5 ค่าโมเมนตัม 0.9 และโหนดของชั้นซ่อนคือ 50 และวิธีสุดท้ายคือ วิธีซัพพอร์ตเวกเตอร์แมชชีน โดยใช้อัลกอริทึม Normalized Poly Kernel, Polynomial Kernel, REF Kernel และ Puk ซึ่งอัลกอริทึมที่มีประสิทธิภาพดีที่สุดของวิธีซัพพอร์ตเวกเตอร์แมชชีนในการทำวิจัยครั้งนี้ คือ อัลกอริทึม Normalized Poly Kernel ซึ่งงานวิจัยนี้ได้ใช้หลักของการทำเหมืองข้อมูลมาใช้ในการจำแนกประเภทข้อมูลและได้ทำการเปรียบเทียบในแต่ละวิธีอีกครั้ง โดยจะใช้การเปรียบเทียบค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุลของการเป็นโรคความดันโลหิตสูง ซึ่งวิธีที่มีค่าความถูกต้องสูงที่สุด คือ วิธีโครงข่ายประสาทเทียม ซึ่งใช้อัลกอริทึมชนิด Multilayer Perceptron วิธีที่ให้ค่าความแม่นยำสูงที่สุด คือ วิธีความใกล้เคียงกันมากที่สุด โดยใช้อัลกอริทึม LWL วิธีที่ให้ค่าความระลึกสูงที่สุด คือ วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ ซึ่งใช้อัลกอริทึมชนิด LMT วิธีโครงข่ายประสาทเทียม ซึ่งใช้อัลกอริทึม ชนิด Multilayer Perceptron และวิธีซัพพอร์ตเวกเตอร์แมชชีน ซึ่งใช้อัลกอริทึมชนิด Normalized Poly Kernel โดยให้ค่าความระลึกเท่ากัน และวิธีที่ให้ค่าความถ่วงดุลสูงที่สุด คือ วิธีโครงข่ายประสาทเทียม ซึ่งใช้อัลกอริทึม ชนิด Multilayer Perceptron ส่วนการเปรียบเทียบเทคนิควิธีที่มีผลการทำนายที่ดีที่สุดของวิธีการจำแนกประเภทของการเป็นโรคความดันโลหิตสูง คือ วิธีโครงข่ายประสาทเทียม ซึ่งใช้อัลกอริทึมชนิด Multilayer Perceptron

เพราะมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย อยู่ในระดับค่อนข้างต่ำที่สุด รองลงมา คือวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ ซึ่งใช้อัลกอริทึมชนิด LMT วิธีความใกล้เคียงกันมากที่สุด ซึ่งใช้อัลกอริทึมชนิด LWL และวิธีซัพพอร์ตเวกเตอร์แมชชีน ซึ่งใช้อัลกอริทึมชนิด NormalizedPoly Kernel ตามลำดับ ซึ่งสามารถสรุปผลได้ดังนี้

จากผลการเปรียบเทียบพบว่า วิธีการจำแนกประเภทที่มีประสิทธิภาพที่ดีที่สุดสำหรับข้อมูลการเป็นโรคความดันโลหิตสูง และวิธีการจำแนกประเภทที่มีผลการทำนายดีที่สุดสำหรับข้อมูลโรคความดันโลหิตสูง คือ วิธีโครงข่ายประสาทเทียม โดยใช้อัลกอริทึมชนิด Multilayer Perceptron โดยมีค่า อัตราการเรียนรู้ 0.5 ค่าโมเมนตัม 0.9 และโหนดของชั้นซ่อนคือ 50 เพราะมีค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุล สูงที่สุด และมีค่าความคลาดเคลื่อนเคลื่อนกำลังสองเฉลี่ยต่ำที่สุด

5.2 ข้อเสนอแนะ

1)ตัวแปรที่นำมาใช้วิเคราะห์ในงานวิจัยนี้เป็นเพียงส่วนหนึ่งของการเกิดโรคความดันโลหิตสูงเท่านั้น เพื่อให้การทำนายมีประสิทธิภาพมากขึ้น ควรเพิ่มตัวแปรที่เกี่ยวข้องอื่นๆอีก

2)เพื่อให้ผลสรุปครอบคลุมกว้างขวางขึ้น ควรจะมีการศึกษาวิธีอื่นๆ ที่เป็นเทคนิคการทำเหมืองข้อมูลในด้านวิธีการจำแนกประเภทเหมือนกัน เช่น วิธีนาอีฟเบย์ (Native Bayes Method) และ โครงข่ายความเชื่อของเบร์เซี่ยน (Bayesian Belief Networks)

3) เพื่อให้ได้ข้อสรุปของผลการวิเคราะห์ข้อมูลที่มีความสมบูรณ์มากขึ้น ดังนั้นเราอาจจะวิเคราะห์ข้อมูลด้วยอัลกอริทึมประเภทอื่นๆ โดยวิธีซัพพอร์ตเวกเตอร์แมชชีน มี Kernel Function อีก 2 แบบ คือ Radial Basis Function Kernel และ Sigmoid Kernel วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ มีอัลกอริทึมที่ใช้ในการจำแนกประเภทอีก เช่น ID3 และวิธีโครงข่ายประสาทเทียม ซึ่งอาจมีการกำหนดค่าอัตราการเรียนรู้ที่ละเอียดขึ้นกว่าเดิม ค่าโมเมนตัมที่ละเอียดขึ้นกว่าเดิมและอาจเพิ่มจำนวนรอบการสอนมากขึ้น

ภาคผนวก ก

รายละเอียดและตัวอย่างข้อมูลที่นำมาวิเคราะห์

1.ข้อมูลของการเป็นโรคความดันโลหิตสูง

ตารางที่ ก-1 คุณลักษณะและรายละเอียดการเป็นโรคความดันโลหิตสูง

คุณลักษณะ	รายละเอียด
1) เพศ (SEX)	ชาย (Male) หญิง (Female)
2) กรรมพันธุ์ (HEREDITY)	มี (Yes) ไม่มี (No) ไม่ทราบ (Unknow)
3) การสูบบุหรี่ (SMOKE)	สูบบุหรี่ (Yes) ไม่สูบบุหรี่ (No) เคยสูบแต่เลิกแล้ว (Ever)
4) การดื่มแอลกอฮอล์ (ALCOHOT)	ดื่มแอลกอฮอล์ (Yes) ไม่ดื่มแอลกอออล์ (No) นานๆครั้ง (Rarely) เคยดื่มแต่เลิกแล้ว (Ever)
5) การออกกำลังกาย (EXERCISE)	ไม่ออกกำลังกาย (Not) ออกกำลังกายน้อยกว่าสัปดาห์ละ 3 ครั้ง (Less) ออกกำลังกายสัปดาห์ละ 3 ครั้ง (Equal) ออกกำลังกายมากกว่าสัปดาห์ละ 3 ครั้งๆละ 30 นาที (More) ออกกำลังกายทุกวันๆละ 30 นาที (Everyday)
6) การรับประทานอาหาร (EATING)	หวาน (Sweet) มัน (Oily) เค็ม (Salt) ไม่ชอบ (No)
7) น้ำหนัก (WEIGHT)	เป็นตัวเลข (Numeric)
8) ส่วนสูง (HEIGHT)	เป็นตัวเลข (Numeric)
9) ค่าดัชนีมวลกาย (BMI)	เป็นตัวเลข (Numeric) คำนวณได้จาก $\mathrm{BMI} = \frac{\mathrm{WEIGHT}}{\mathrm{HEIGHT}^2}$ กก./ม^2

ตารางที่ ก-1 คุณลักษณะและรายละเอียดการเป็นโรคความดันโลหิตสูง (ต่อ)

คุณลักษณะ	รายละเอียด
10) เกณฑ์ดัชนีมวลกาย (BMI MARK)	น้ำหนักน้อยเกินไป (Skinny) น้ำหนักปกติ (Thin) น้ำหนักเกิน (Shapely) อ้วนระดับ 1 (Plump) อ้วนระดับ 2 (Fat)
11) รอบเอว (WAISTLINE)	เป็นตัวเลข (Numeric)
12) เกณฑ์ค่าความดันโลหิต (BLOOD PRESSURE)	ค่าความดันโลหิตอยู่ในเกณฑ์ปกติ (Normal) ค่าความดันโลหิตอยู่ในเกณฑ์มากกว่าปกติ (Hypertension)
13) การเป็นโรคความดันโลหิต (HYPERTENSION)	เป็นโรคความดันโลหิตสูง (Positive) ไม่เป็นโรคความดันโลหิตสูง (Negative)

ตารางที่ ก-2 ตัวอย่างข้อมูลการเป็นโรคความดันโลหิตสูง

No	SEX	RELATIVE	SMOKE	ALCOHOL	EXERCISE	EATING
1	FEMALE	No	No	No	Less	No
2	FEMALE	No	Yes	No	Less	No
3	MALE	No	Yes	Yes	More	Salt
4	FEMALE	No	No	Ever	Not	Sweet
5	FEMALE	No	No	No	Not	Sweet
6	FEMALE	No	No	No	Not	Salt
7	FEMALE	Yes	No	No	Less	Sweet
8	FEMALE	Yes	No	No	Not	Salt
9	MALE	Yes	No	No	Equal	Sweet
10	MALE	Yes	Yes	Yes	Not	No
11	FEMALE	Yes	No	No	Everyday	Sweet
12	FEMALE	Yes	No	No	Everyday	Oily
13	MALE	Yes	Yes	No	Less	Oily
14	FEMALE	Yes	No	No	Less	Oily
15	FEMALE	Yes	No	No	Equal	Salt
16	FEMALE	Yes	Yes	Rarely	Equal	Sweet
17	MALE	Yes	Yes	Rarely	Everyday	Salt
18	MALE	Yes	Yes	Rarely	Less	Oily
19	MALE	Yes	Yes	Yes	Not	No
20	MALE	No	Ever	Ever	Less	Salt
21	FEMALE	Unknow	No	No	Equal	No
22	FEMALE	No	No	No	Not	No
23	FEMALE	No	No	No	Less	No
24	FEMALE	No	No	Yes	Less	No
25	FEMALE	No	No	Ever	Less	No
26	FEMALE	Unknow	No	No	Less	Salt
27	FEMALE	No	No	No	Less	Salt
28	FEMALE	No	Yes	Yes	Equal	Sweet
29	FEMALE	No	No	Yes	Less	Oily
30	FEMALE	No	No	Rarely	Everyday	Salt

ตารางที่ ก-2 ตัวอย่างข้อมูลการเป็นโรคความดันโลหิตสูง (ต่อ)

No	WEIGHT	HIGH	BMI	BMI MARK	WAISTLINE
1	70	1.6	27.3	Plump	111.76
2	60	1.6	23.4	Shapely	80
3	65	1.6	25.4	Plump	85
4	55	1.56	22.6	Thin	77
5	45	1.55	18.7	Thin	77
6	91	1.65	33.4	Fat	106
7	67	1.6	26.2	Plump	94
8	76	1.55	31.6	Fat	108
9	53	1.65	19.5	Thin	75
10	80	1.73	26.7	Plump	96
11	50	1.57	20.3	Thin	84
12	62	1.55	25.8	Plump	81.28
13	57	1.67	20.4	Thin	71
14	54	1.53	23.1	Shapely	79
15	51	1.5	22.7	Thin	74
16	50	1.65	18.4	Skinny	76
17	89	1.78	28.1	Plump	96
18	68	1.61	26.2	Plump	80
19	53	1.7	18.3	Skinny	77
20	58	1.63	21.8	Thin	79
21	64	1.55	26.6	Plump	89
22	50	1.5	22.2	Thin	94
23	57	1.6	22.3	Thin	78
24	60	1.6	23.4	Shapely	82
25	68	1.67	24.4	Shapely	85
26	47	1.5	20.9	Thin	72
27	66	1.68	23.4	Shapely	89
28	64	1.7	22.1	Thin	86
29	54	1.5	24	Shapely	87
30	55	1.65	20.2	Thin	71

ตารางที่ ก-2 ตัวอย่างข้อมูลการเป็นโรคความดันโลหิตสูง (ต่อ)

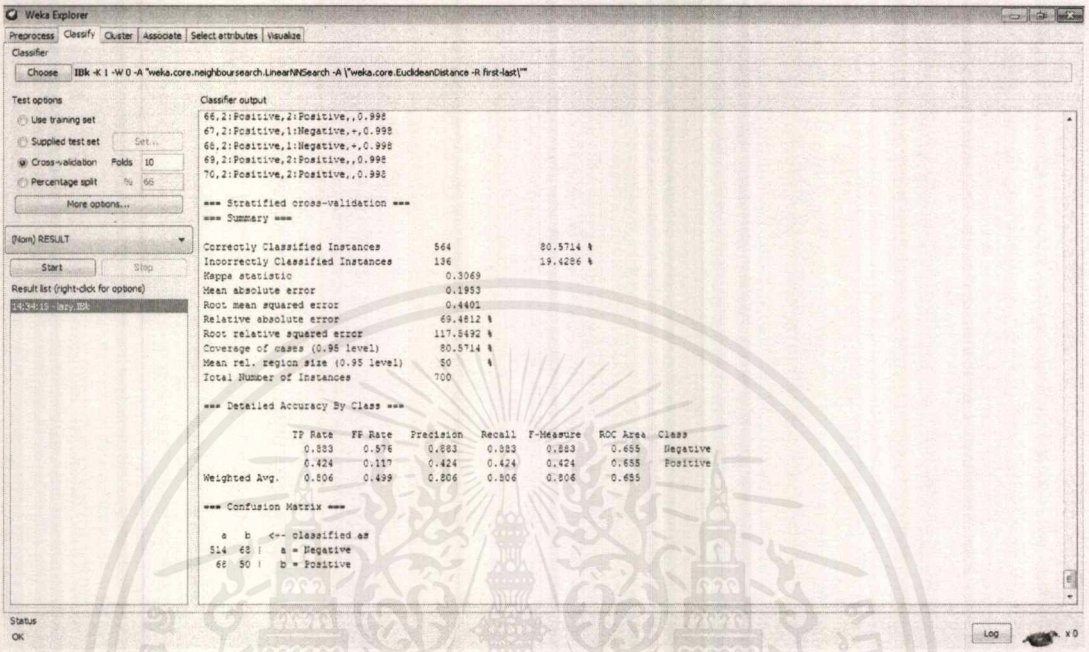
No	BLOOD PRESSURE	HYPERTENSION
1	Normal	Negative
2	Normal	Negative
3	Normal	Negative
4	Hypertension	Negative
5	Normal	Negative
6	Hypertension	Positive
7	Hypertension	Positive
8	Normal	Negative
9	Normal	Negative
10	Normal	Negative
11	Normal	Positive
12	Hypertension	Positive
13	Normal	Negative
14	Normal	Negative
15	Normal	Negative
16	Hypertension	Negative
17	Normal	Negative
18	Normal	Negative
19	Normal	Negative
20	Normal	Negative
21	Normal	Negative
22	Normal	Negative
23	Normal	Negative
24	Hypertension	Positive
25	Normal	Negative
26	Normal	Negative
27	Normal	Positive
28	Hypertension	Negative
29	Normal	Positive
30	Normal	Negative

ภาคผนวก ข

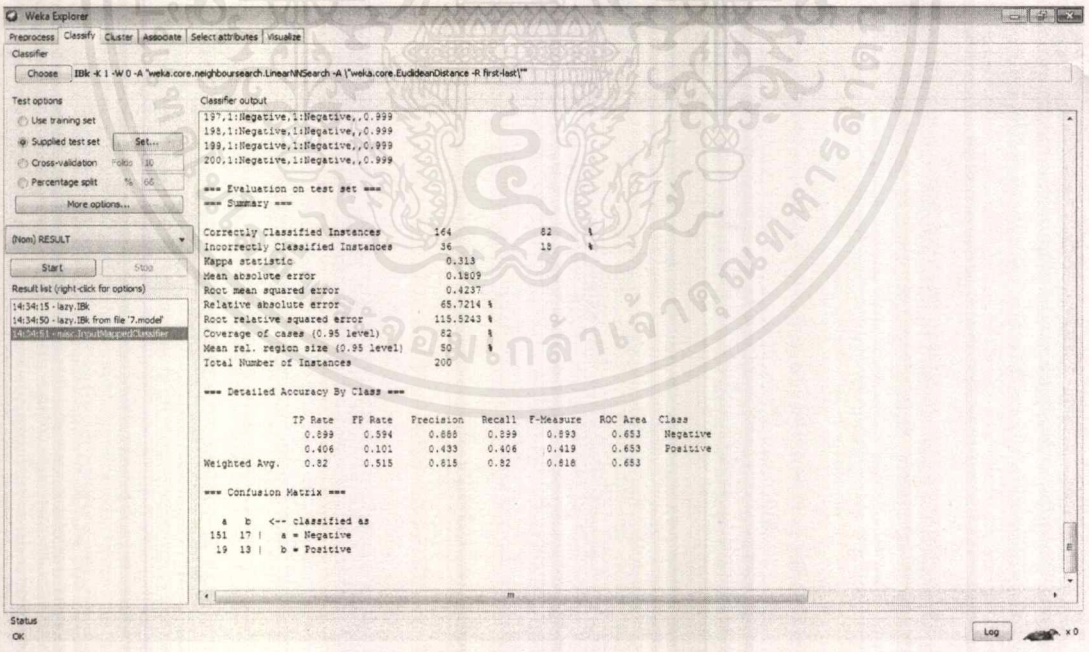
การวิเคราะห์ข้อมูล

1.วิธีความใกล้เคียงกันมากที่สุด

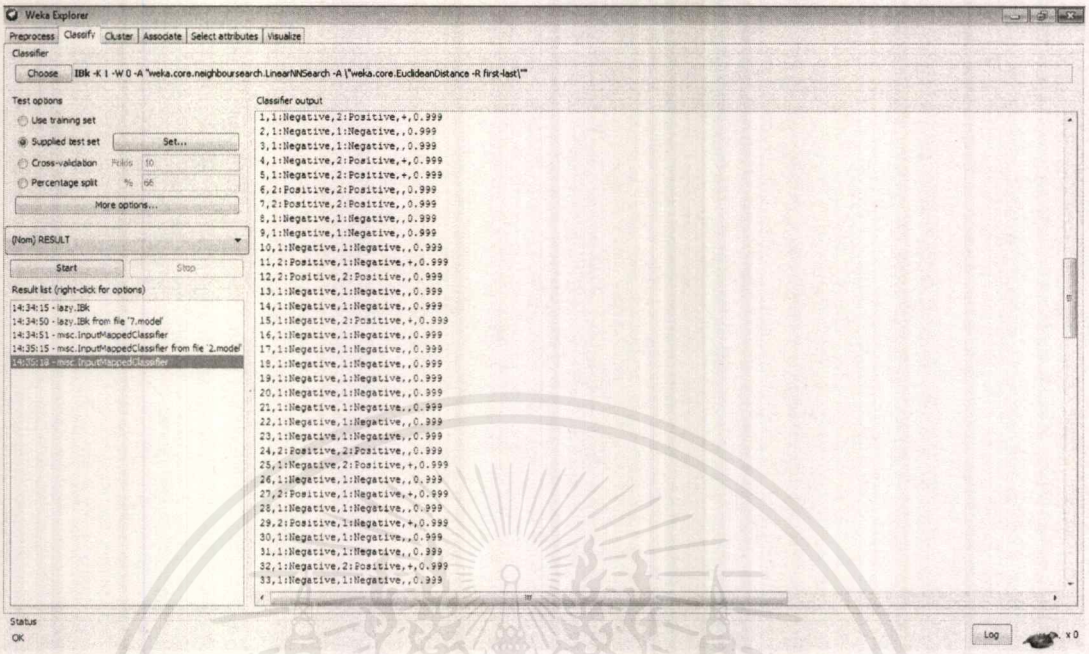
1.1 อัลกอริทึม IBk



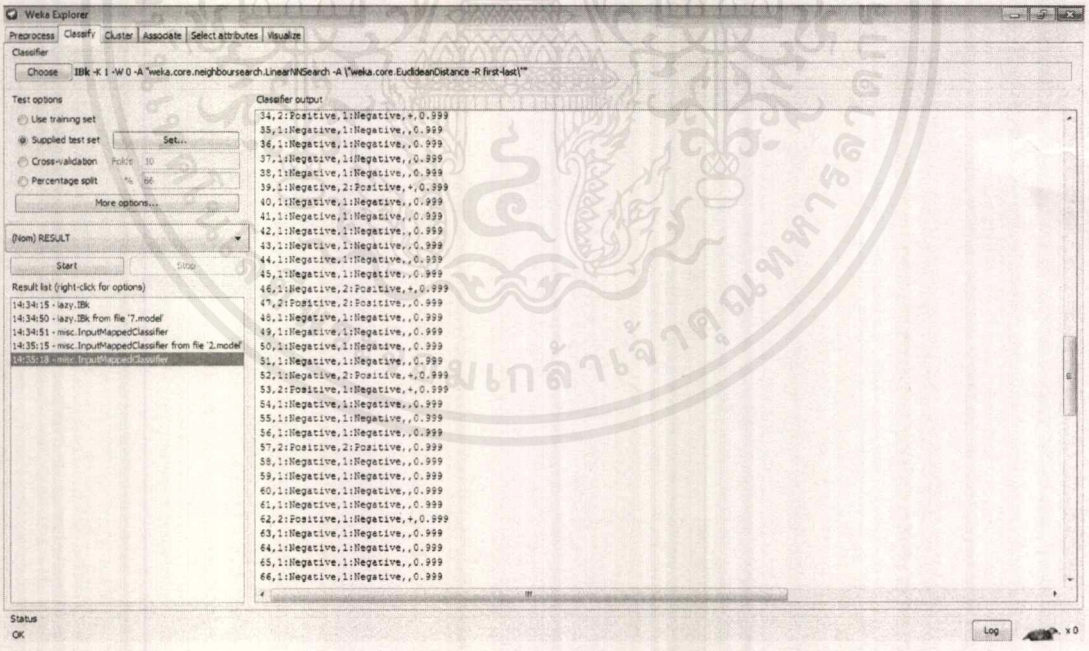
รูปที่ ข-1 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการสร้างตัวแบบโดยอัลกอริทึม IBk



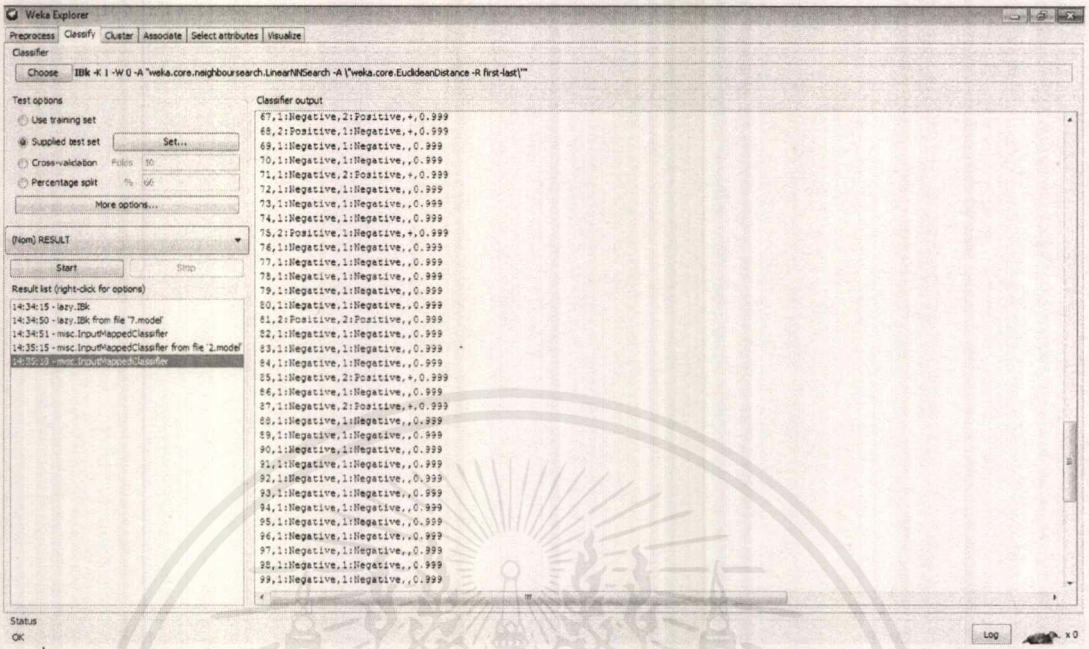
รูปที่ ข-2 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทดสอบตัวแบบโดยอัลกอริทึม IBk



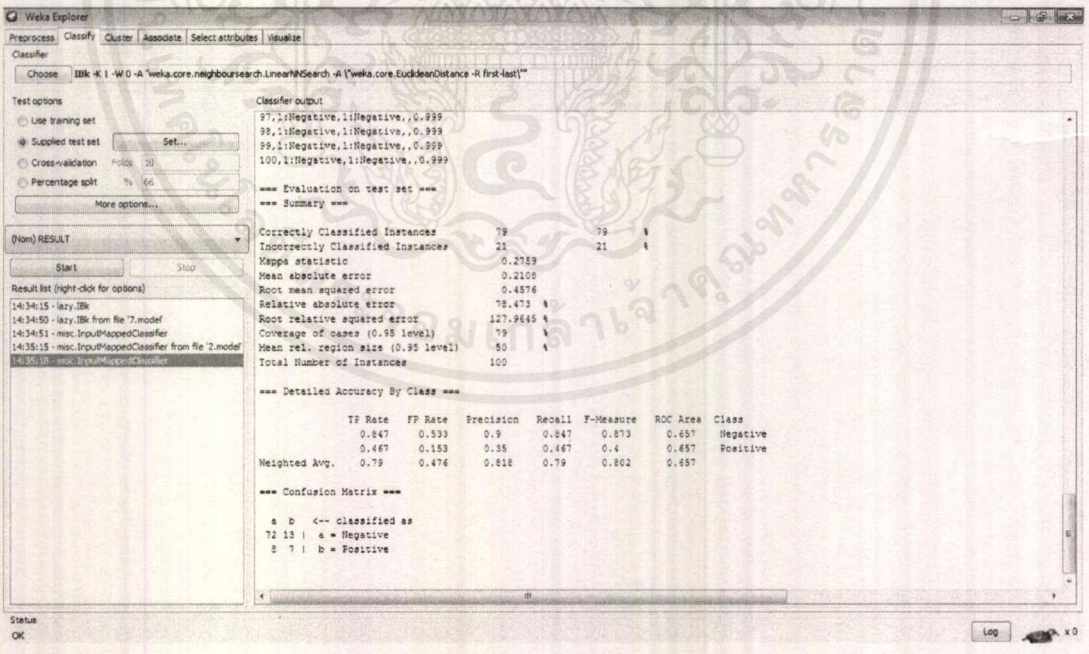
รูปที่ ข-3 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม IBk



รูปที่ ข-4 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม IBk

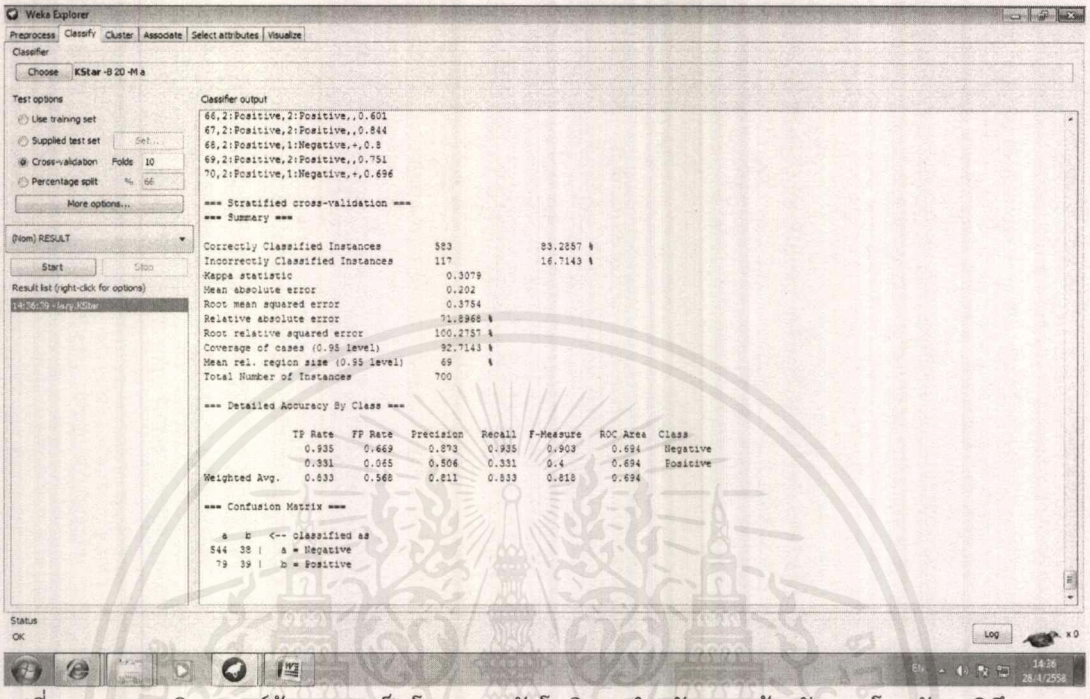


รูปที่ ข-5 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม IBk

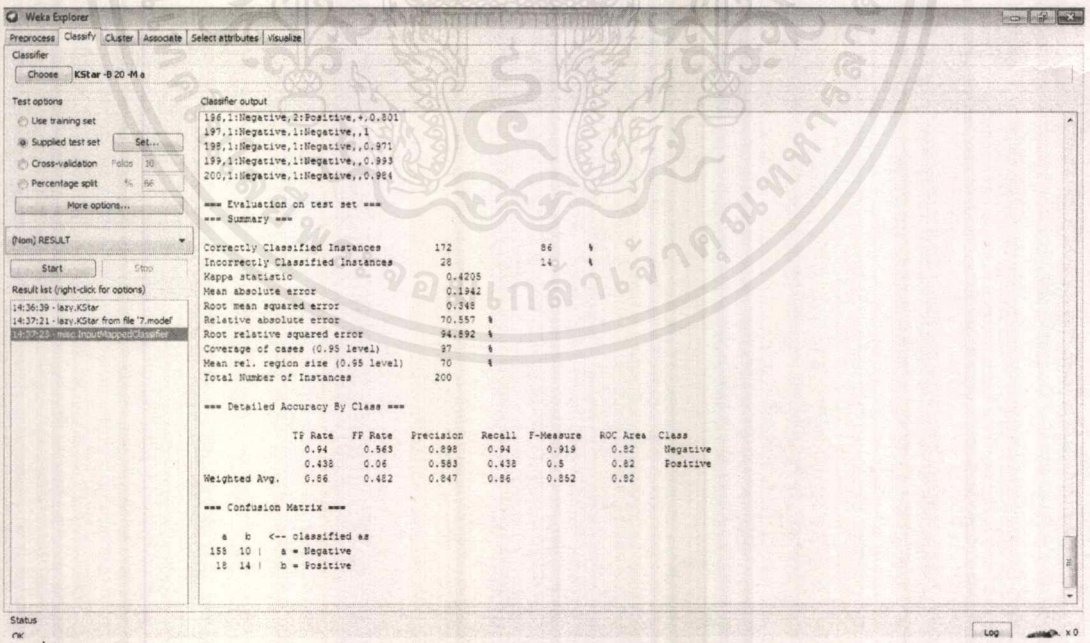


รูปที่ ข-6 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม IBk

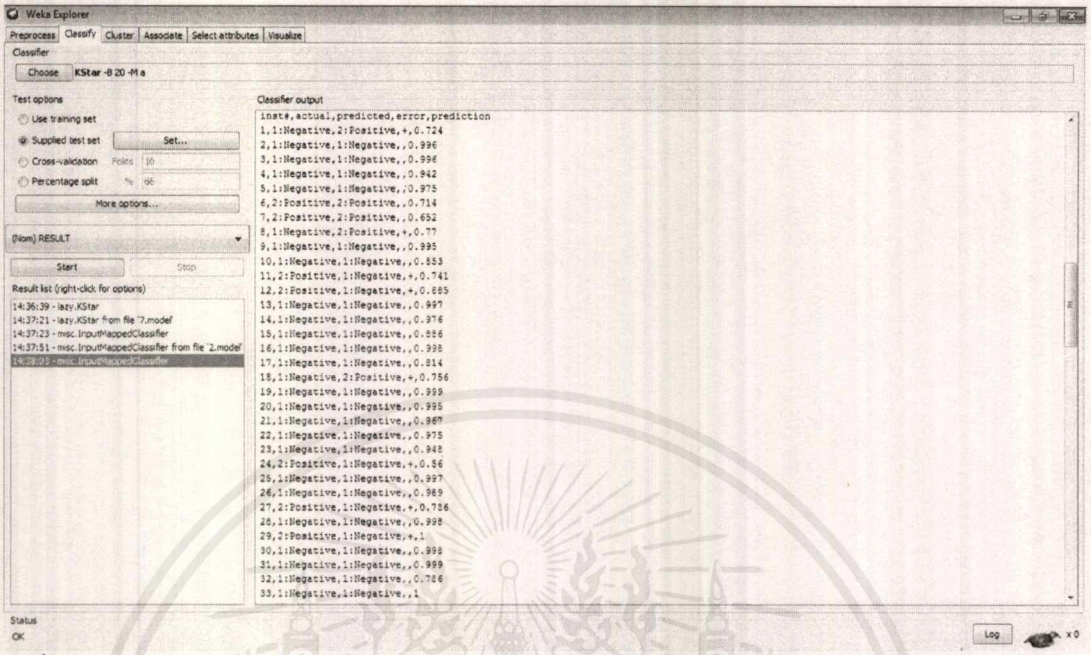
1.2 อัลกอริทึม KStar



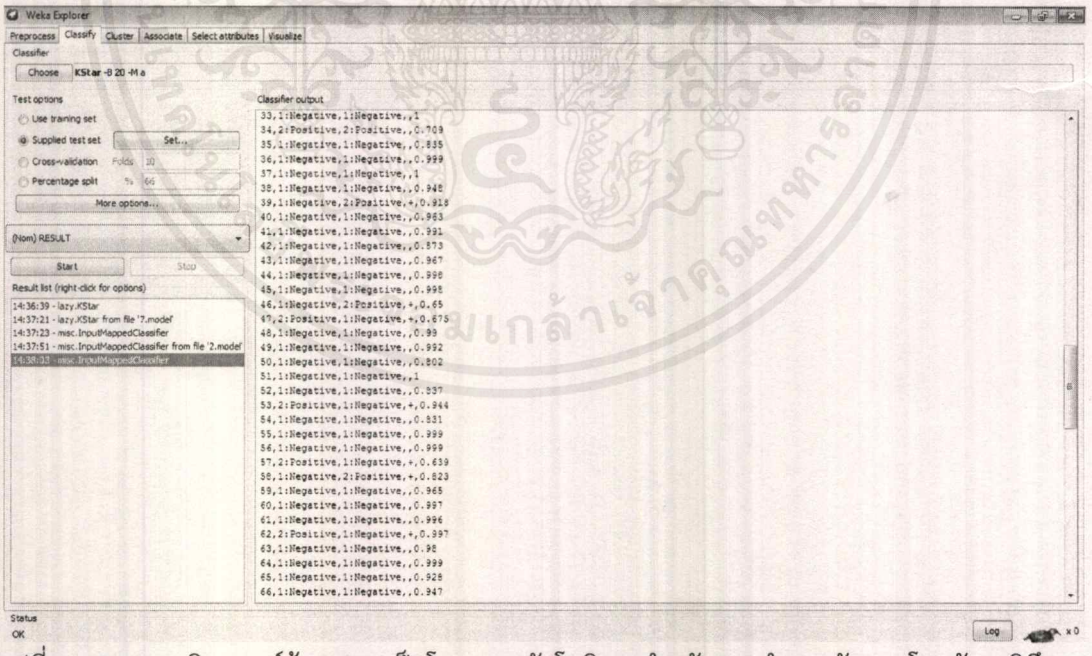
รูปที่ ข-7 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการสร้างตัวแบบโดยอัลกอริทึม KStar



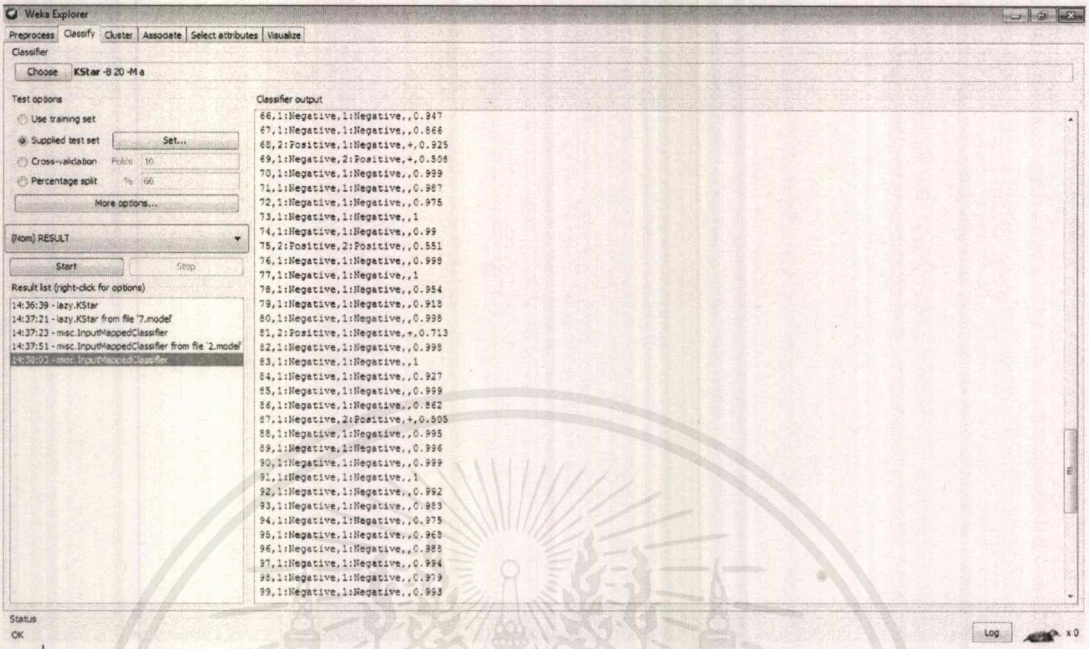
รูปที่ ข-8 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทดสอบตัวแบบโดยอัลกอริทึม KStar



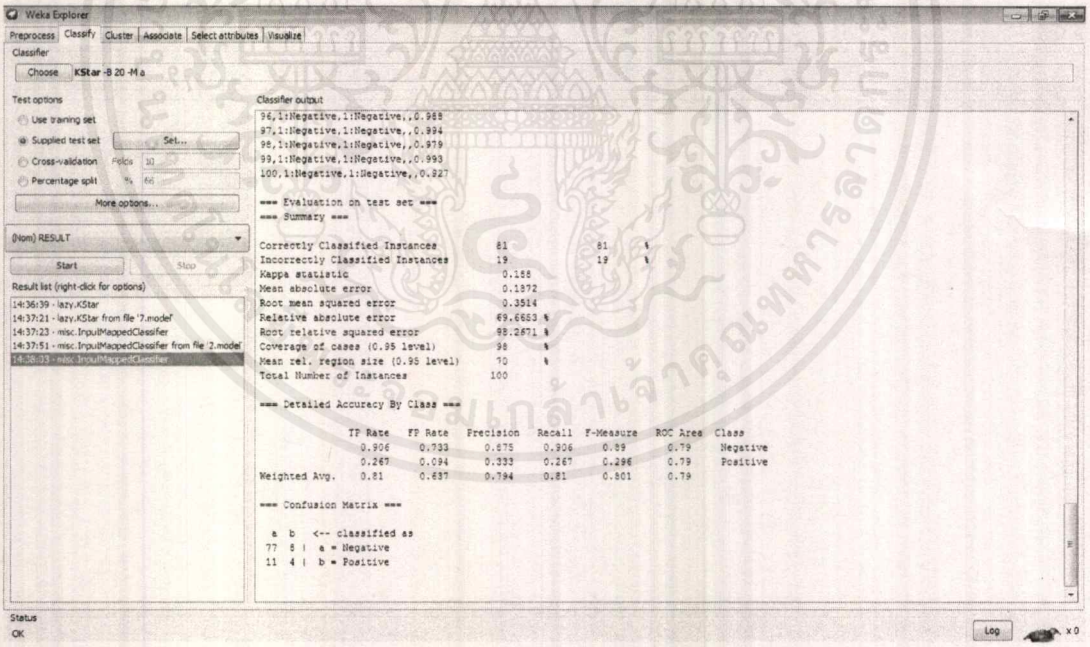
รูปที่ ข-9 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม KStar



รูปที่ ข-10 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม KStar

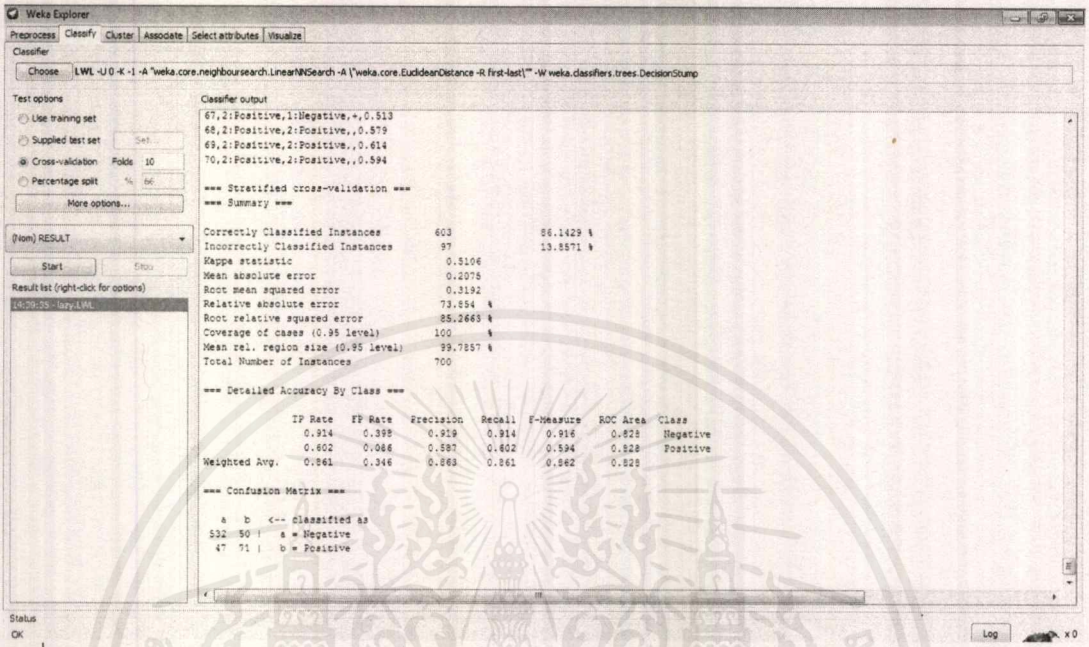


รูปที่ ข-11 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม KStar

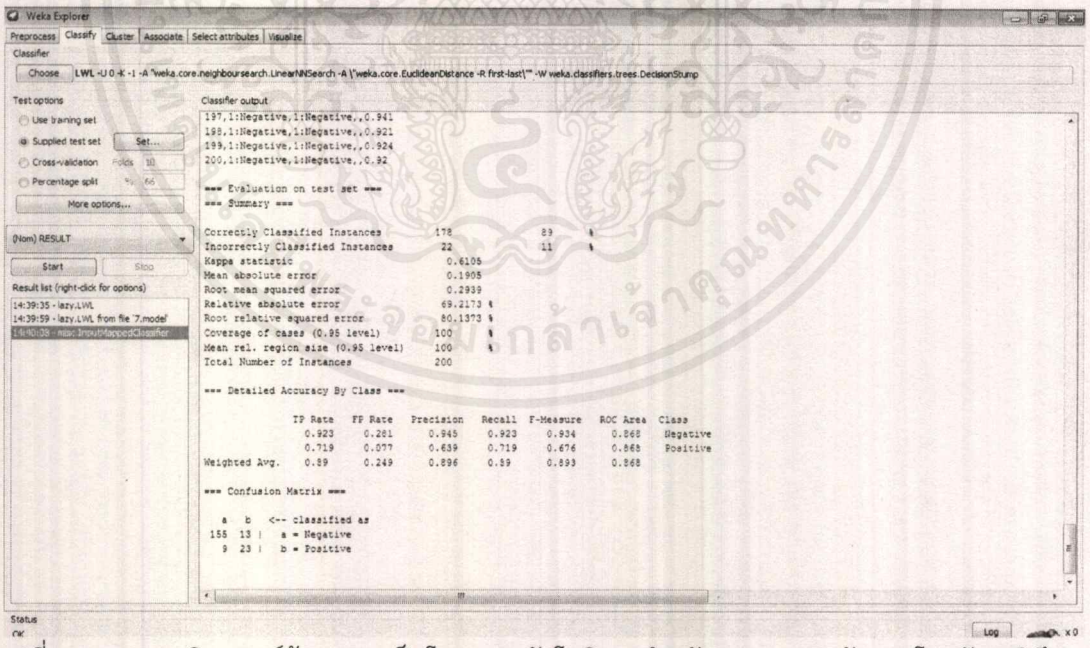


รูปที่ ข-12 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม KStar

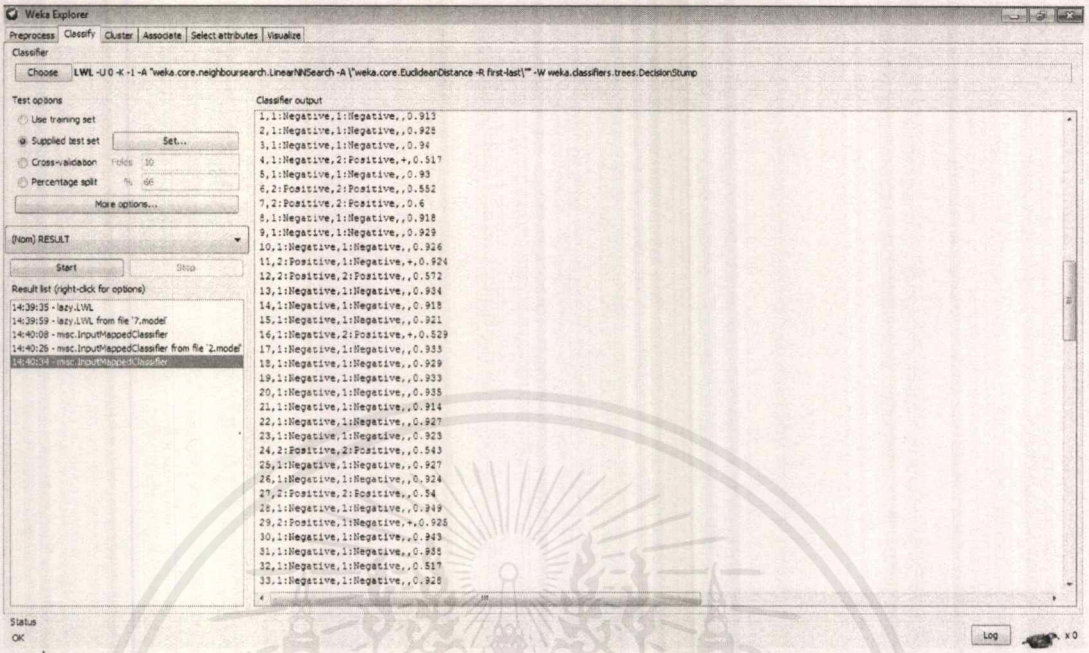
1.3 อัลกอริทึม LWL



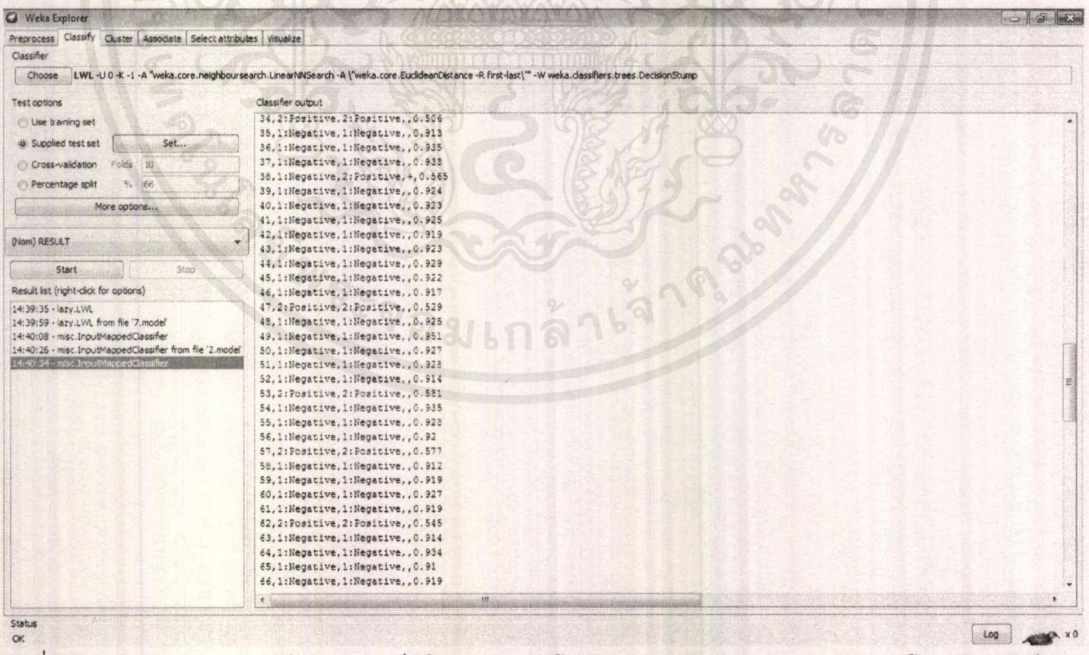
รูปที่ ข-13 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการสร้างตัวแบบโดยอัลกอริทึม LWL



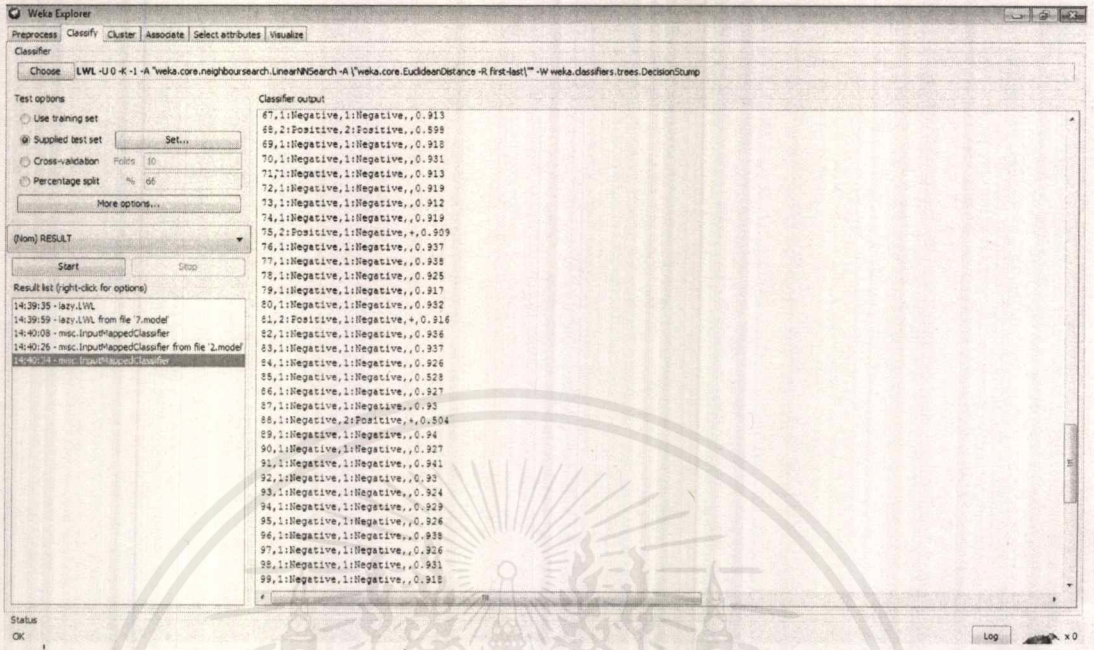
รูปที่ ข-14 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทดสอบตัวแบบโดยอัลกอริทึม LWL



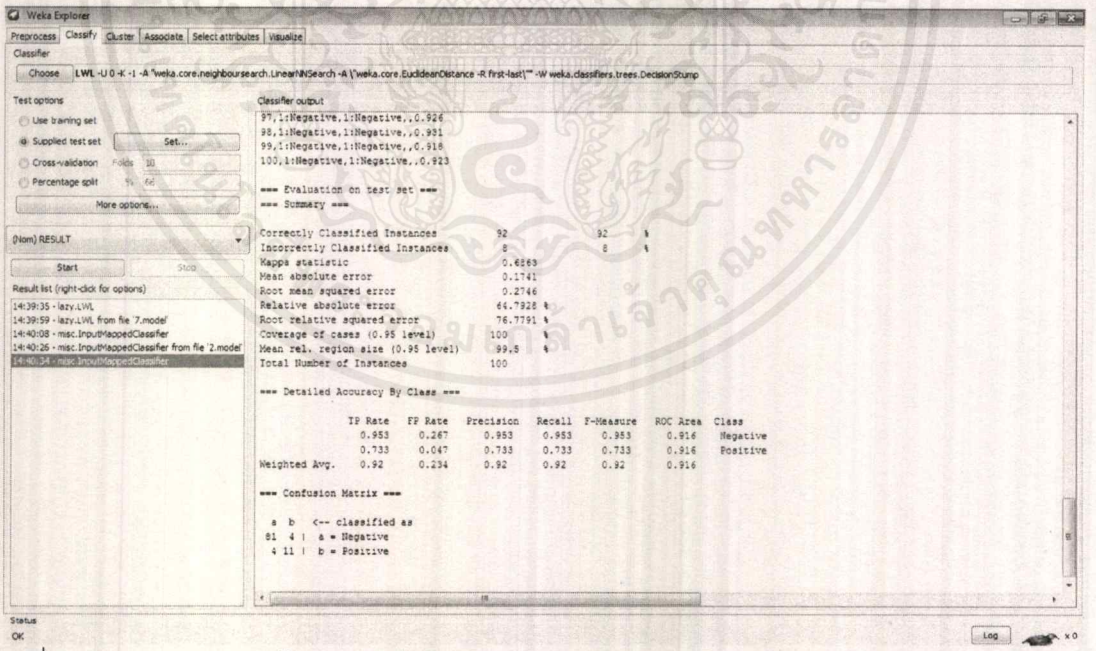
รูปที่ ข-15 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม LWL



รูปที่ ข-16 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม LWL



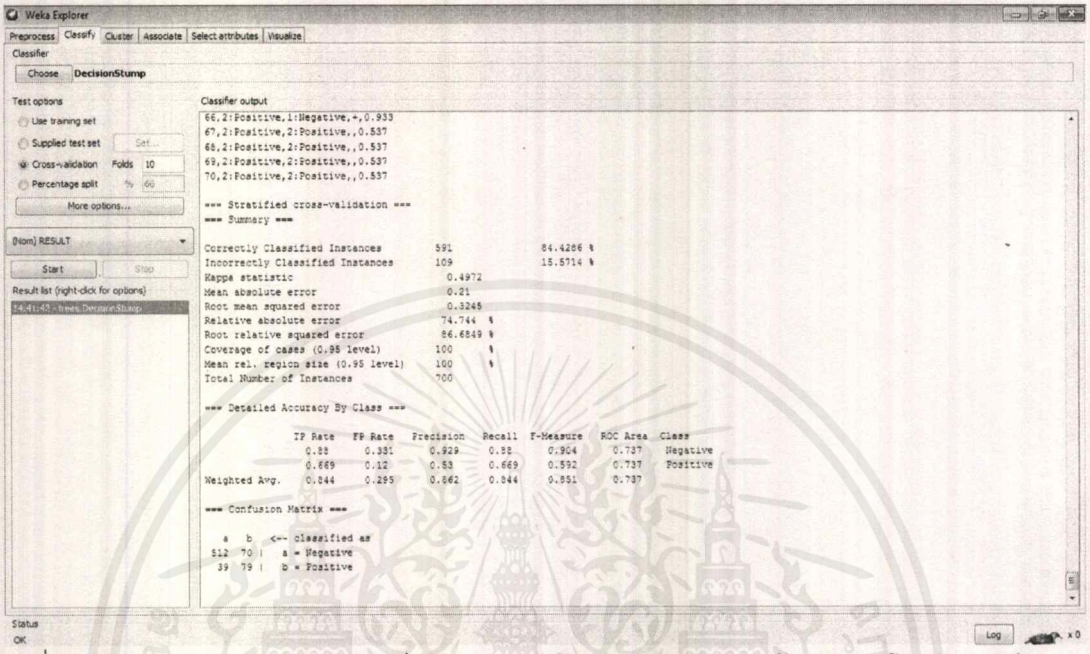
รูปที่ ข-17 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม LWL



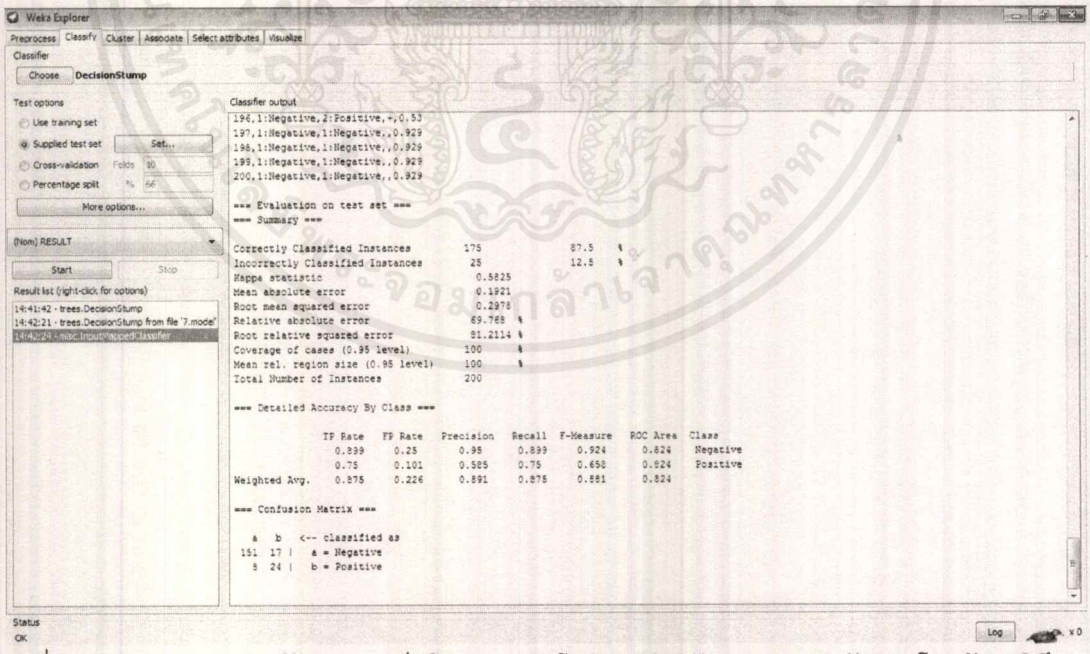
รูปที่ ข-18 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม LWL

2.วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ

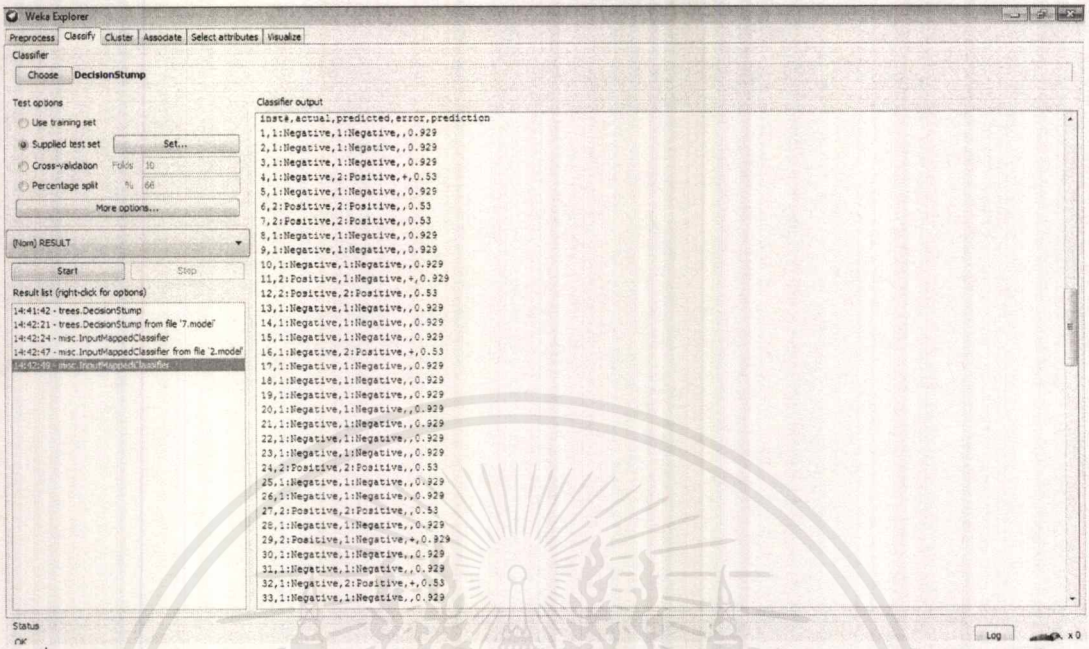
2.1 อัลกอริทึม Decision Stump



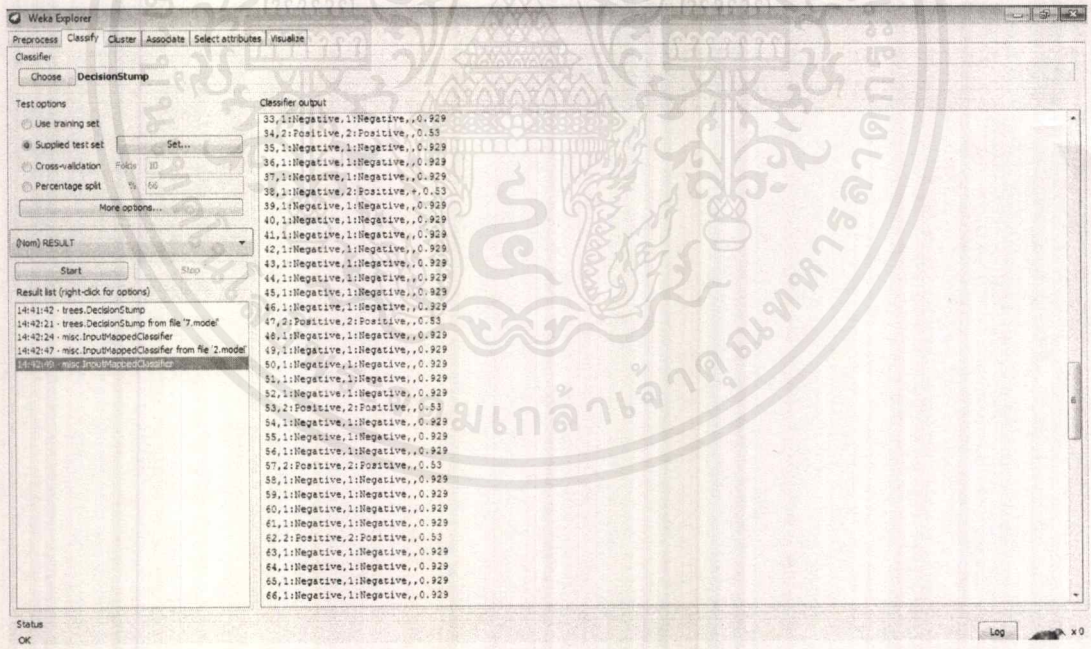
รูปที่ ข-19 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการสร้างตัวแบบโดยอัลกอริทึม Decision Stump



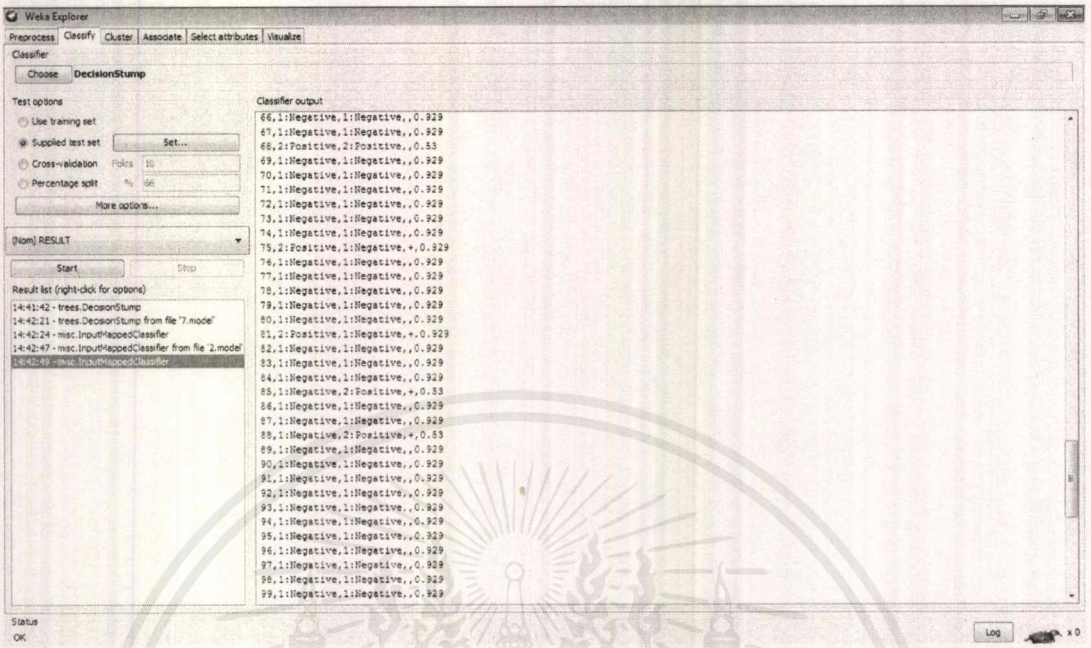
รูปที่ ข-20 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทดสอบตัวแบบโดยอัลกอริทึม Decision Stump



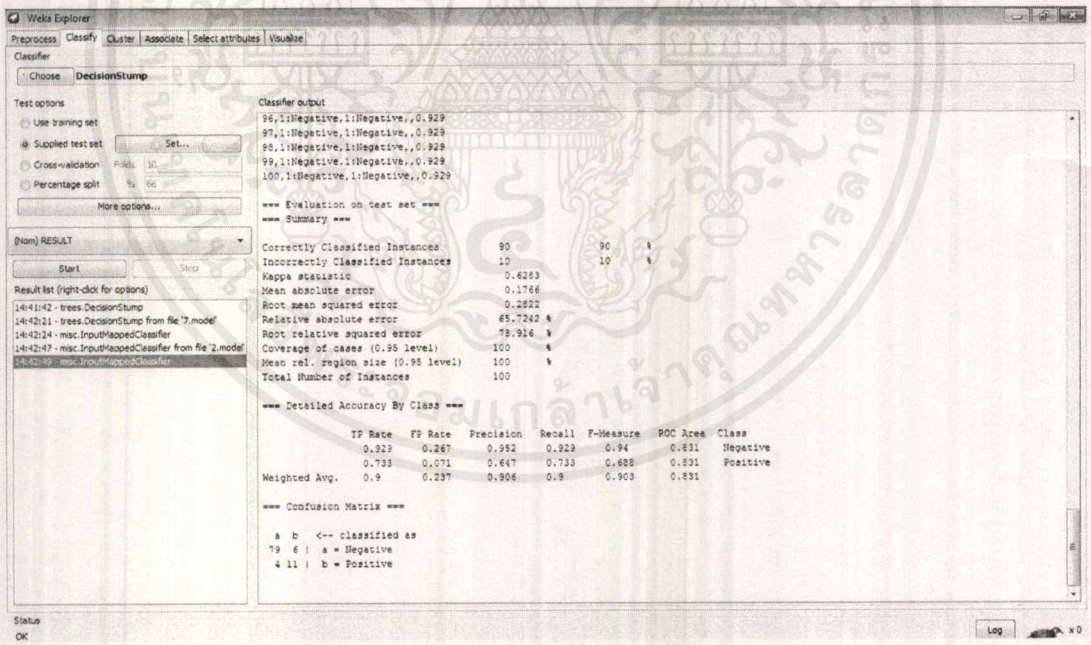
รูปที่ ข-21 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม Decision Stump



รูปที่ ข-22 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม Decision Stump

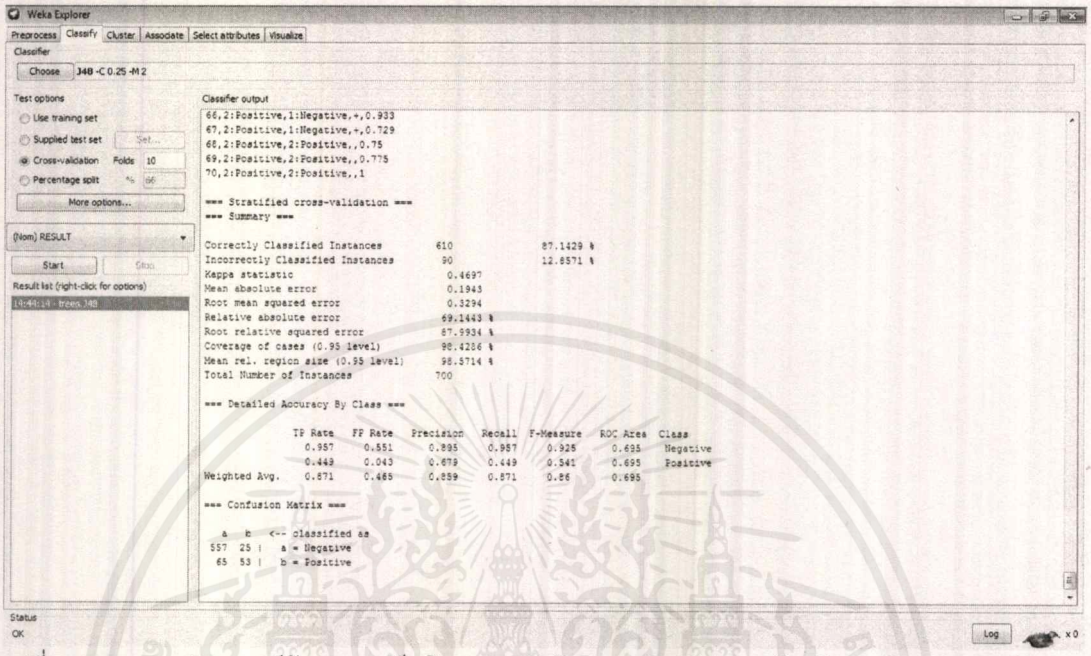


รูปที่ ข-23 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม Decision Stump

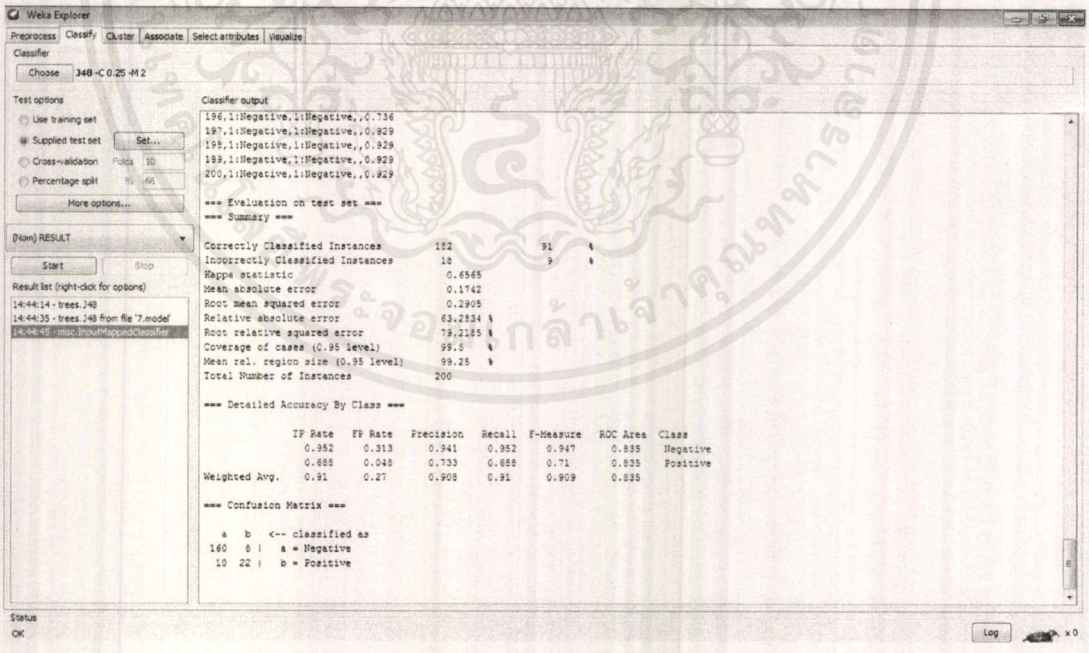


รูปที่ ข-24 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม Decision Stump

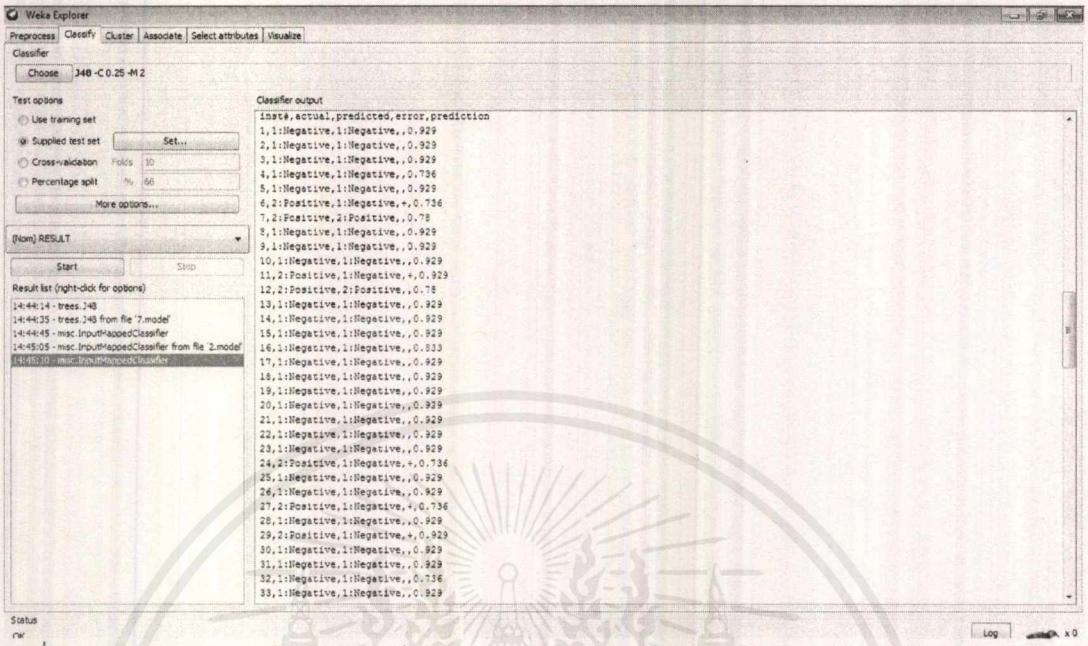
2.2 อัลกอริทึม J48



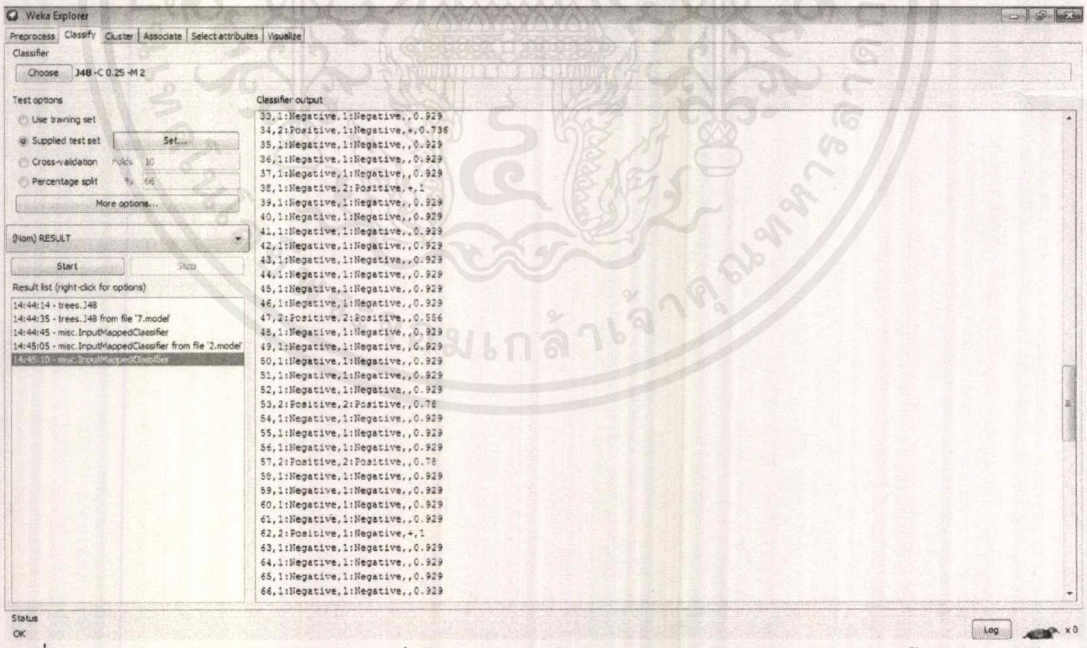
รูปที่ ข-25 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการสร้างตัวแบบโดยอัลกอริทึม J48



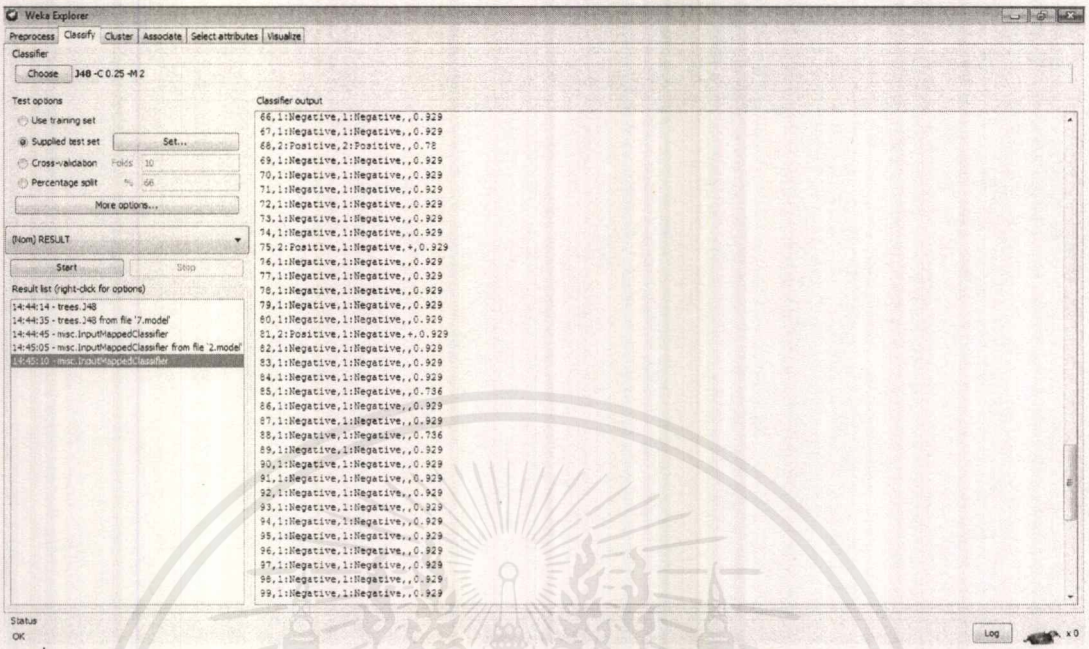
รูปที่ ข-26 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทดสอบตัวแบบโดยอัลกอริทึม J48



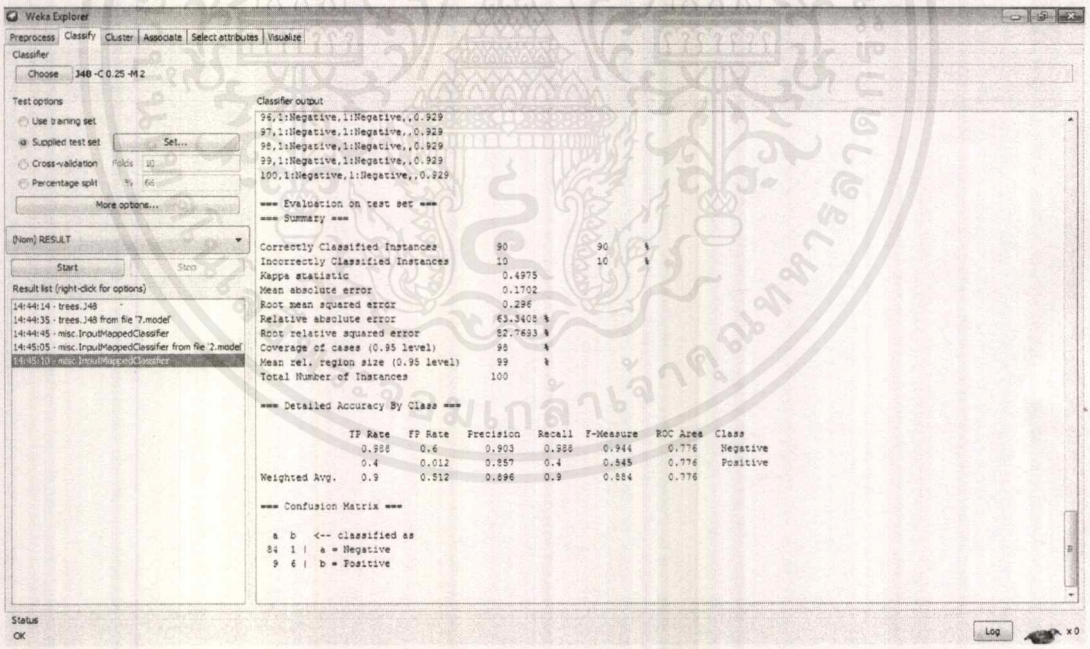
รูปที่ ข-27 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม J48



รูปที่ ข-28 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม J48

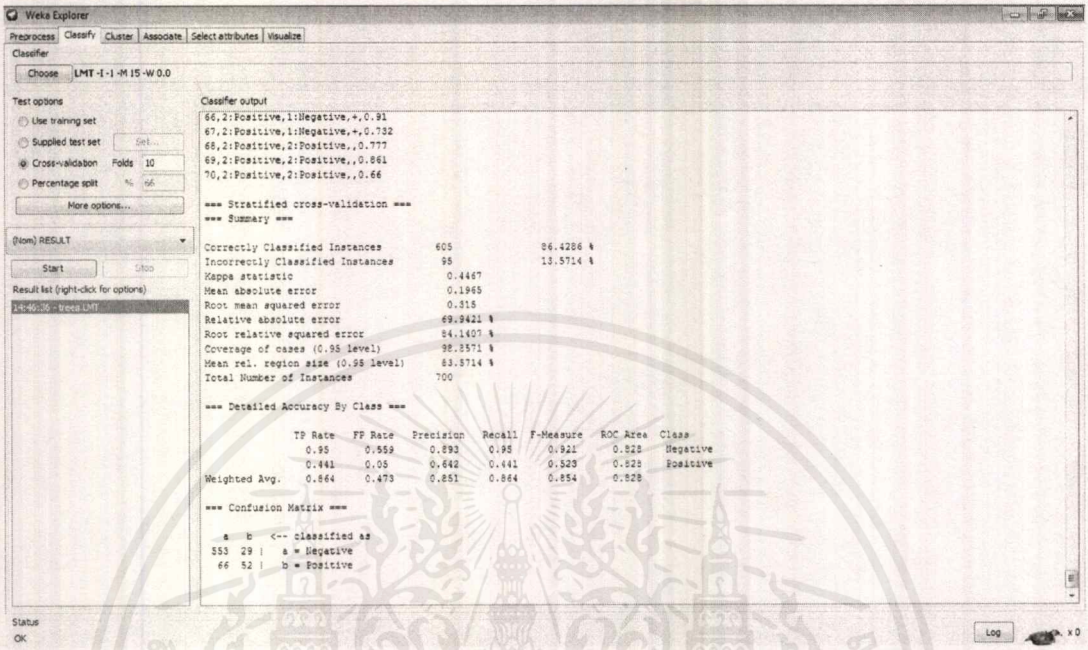


รูปที่ ข-29 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม J48

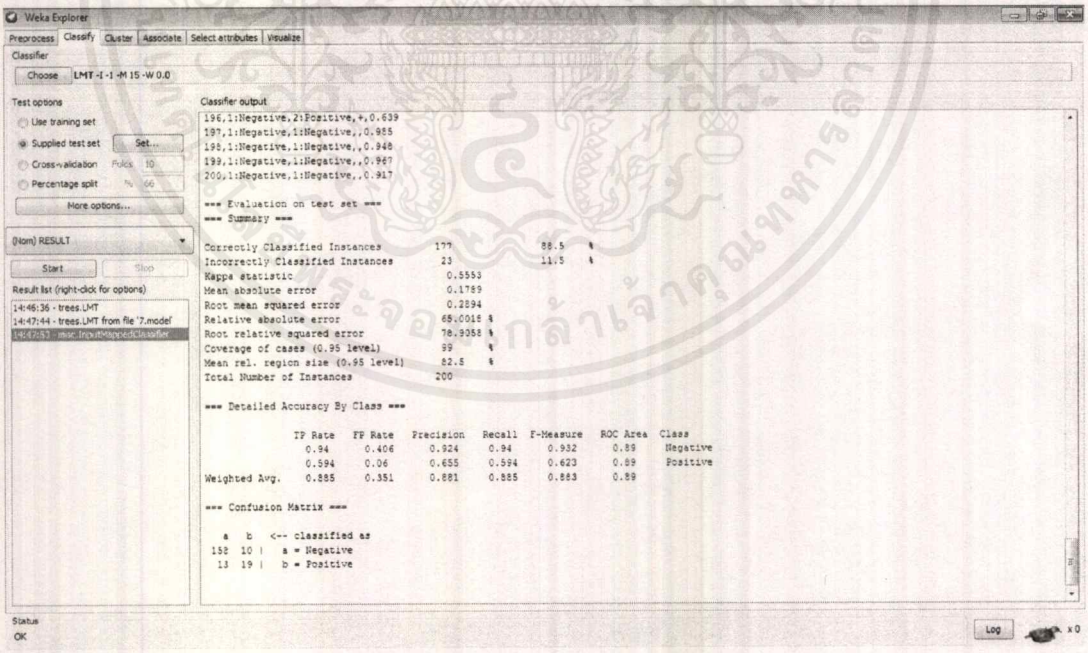


รูปที่ ข-30 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม J48

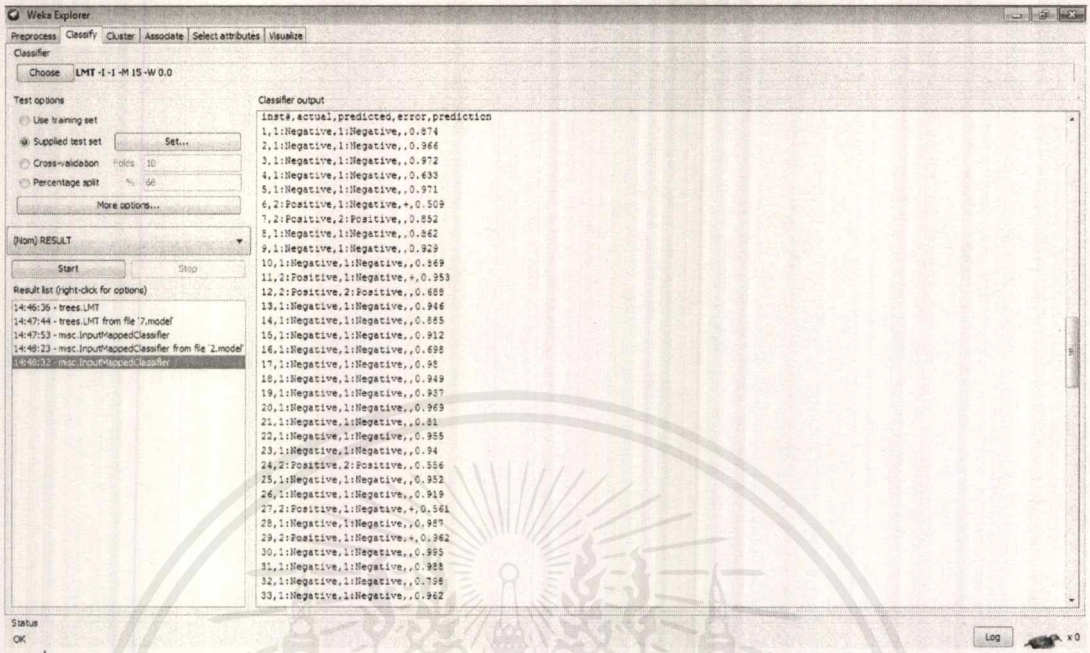
2.3 อัลกอริทึม LMT



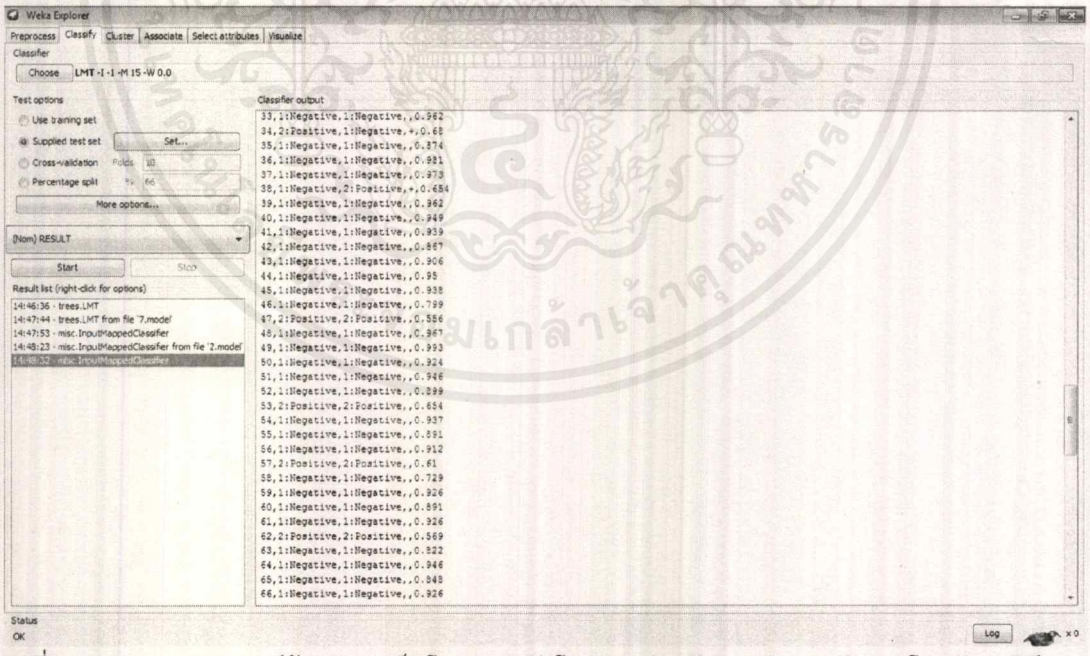
รูปที่ ข-31 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการสร้างตัวแบบโดยอัลกอริทึม LMT



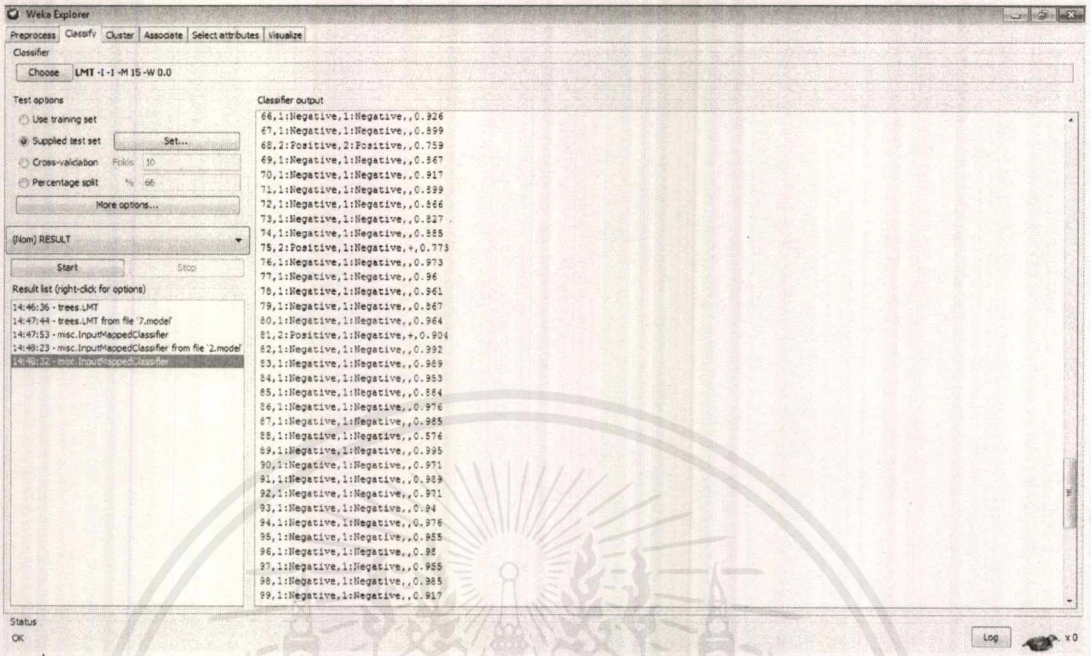
รูปที่ ข-32 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทดสอบตัวแบบโดยอัลกอริทึม LMT



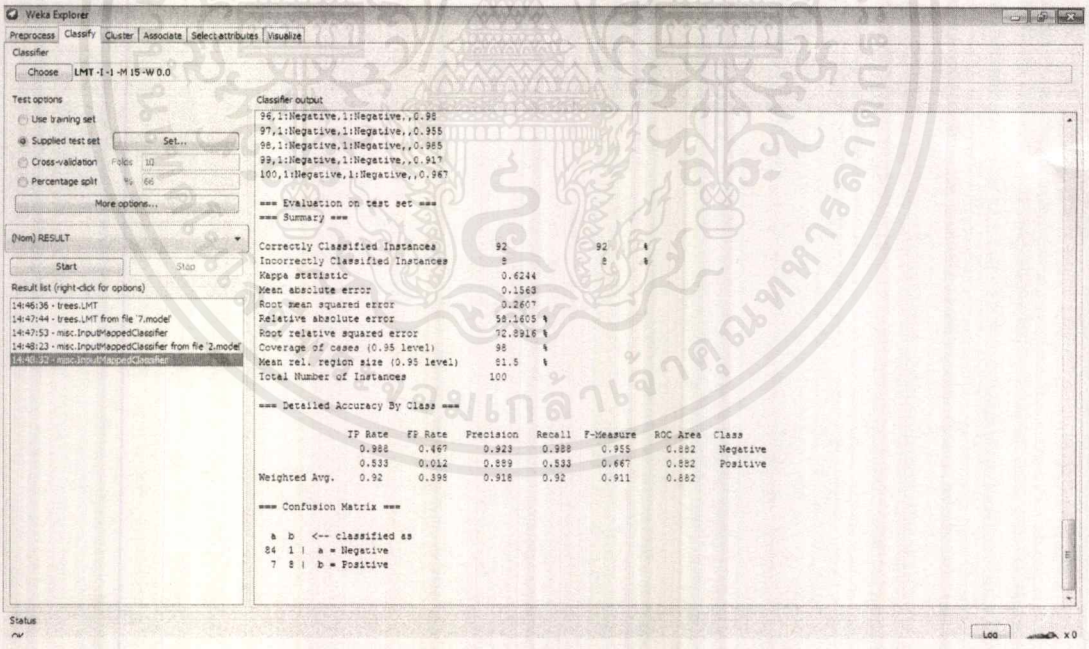
รูปที่ ข-33 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม LMT



รูปที่ ข-34 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม LMT

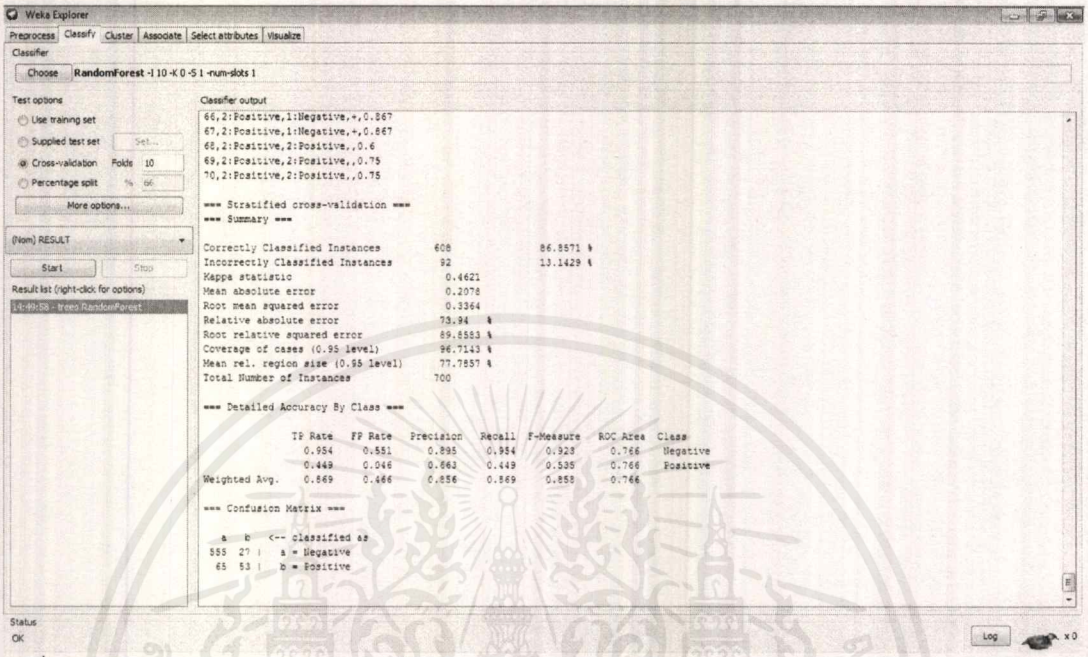


รูปที่ ข-35 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม LMT

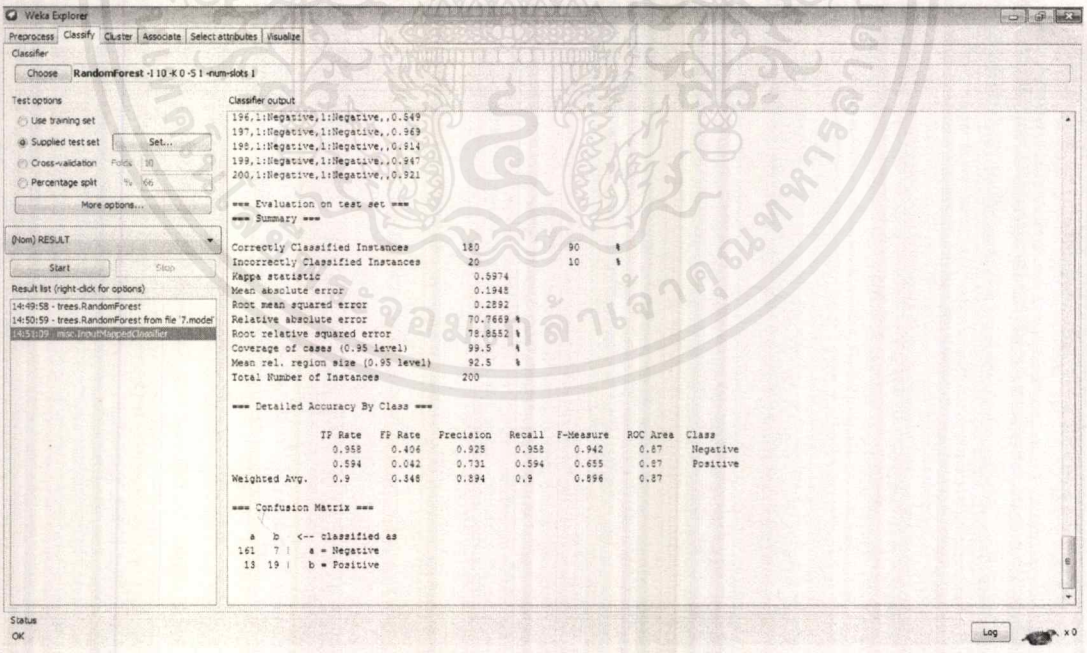


รูปที่ ข-36 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม LMT

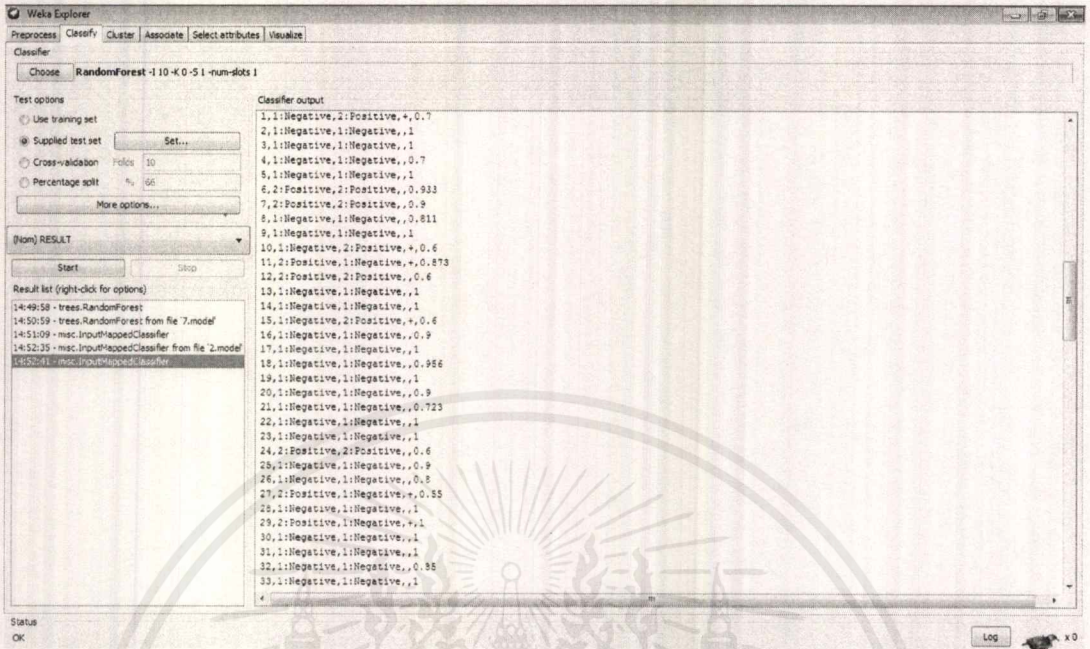
2.4 อัลกอริทึม Random Forest



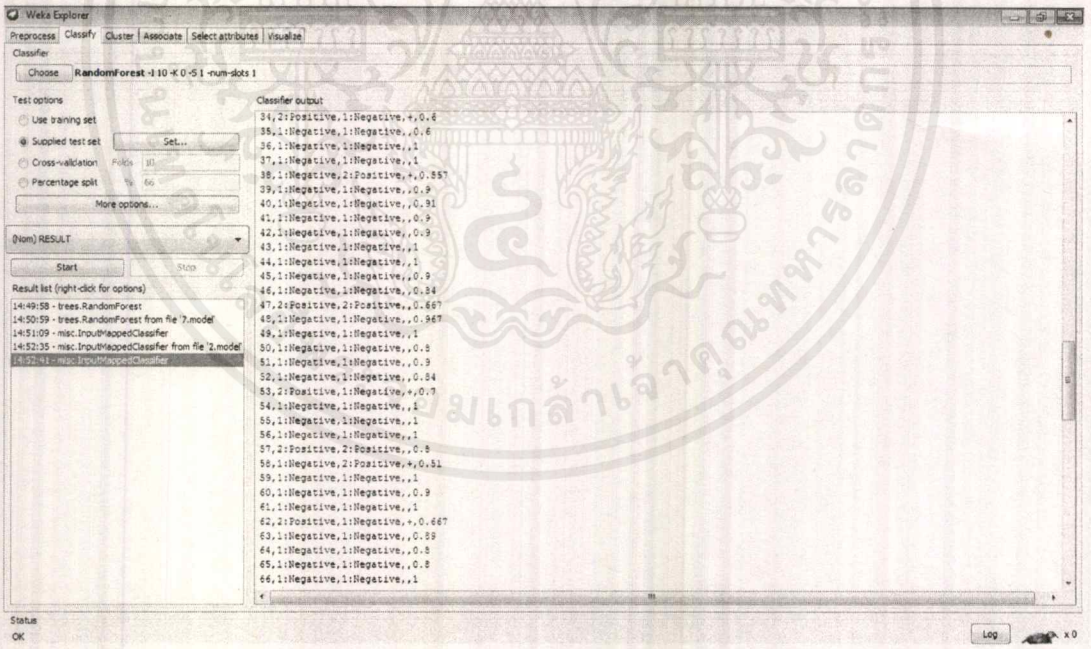
รูปที่ ข-37 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการสร้างตัวแบบโดยอัลกอริทึม Random Forest



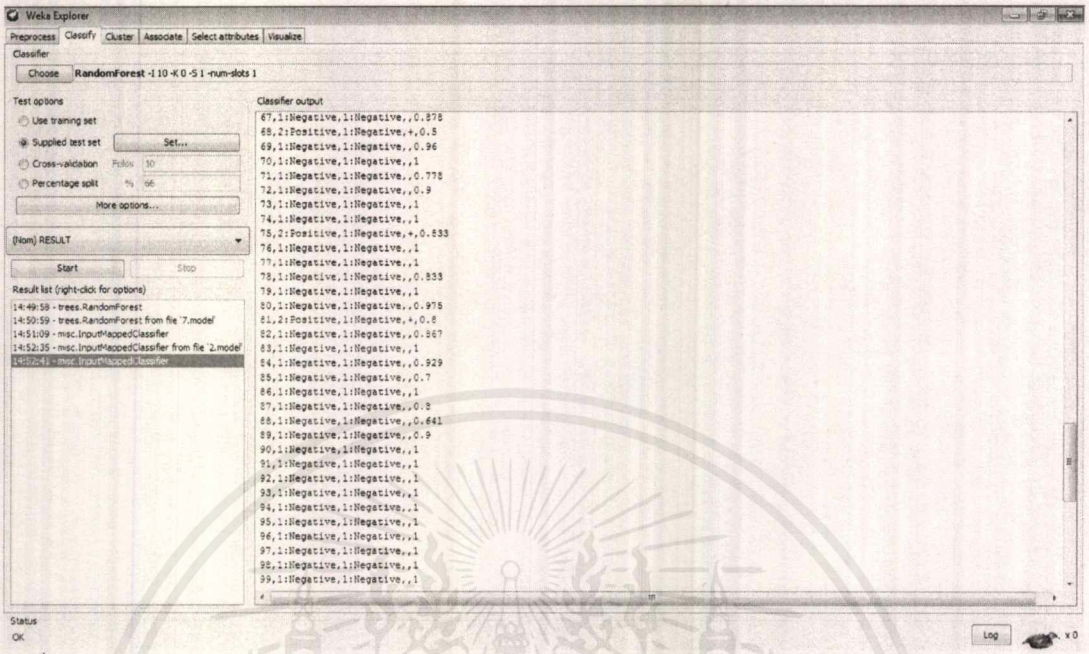
รูปที่ ข-38 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทดสอบตัวแบบโดยอัลกอริทึม Random Forest



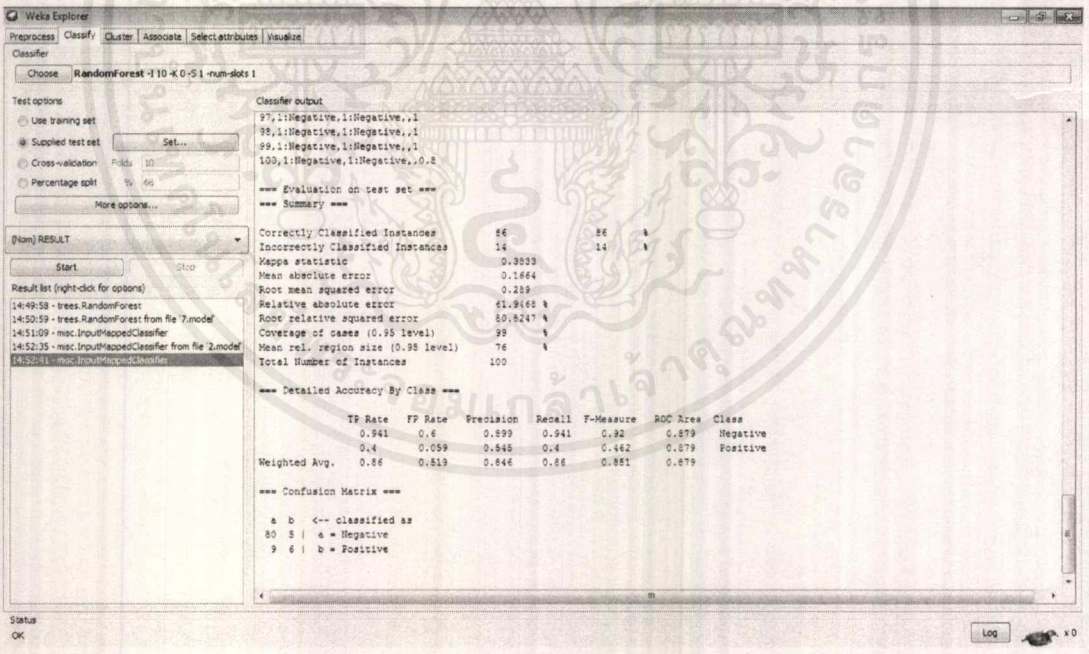
รูปที่ ข-39 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม Random Forest



รูปที่ ข-40 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม Random Forest

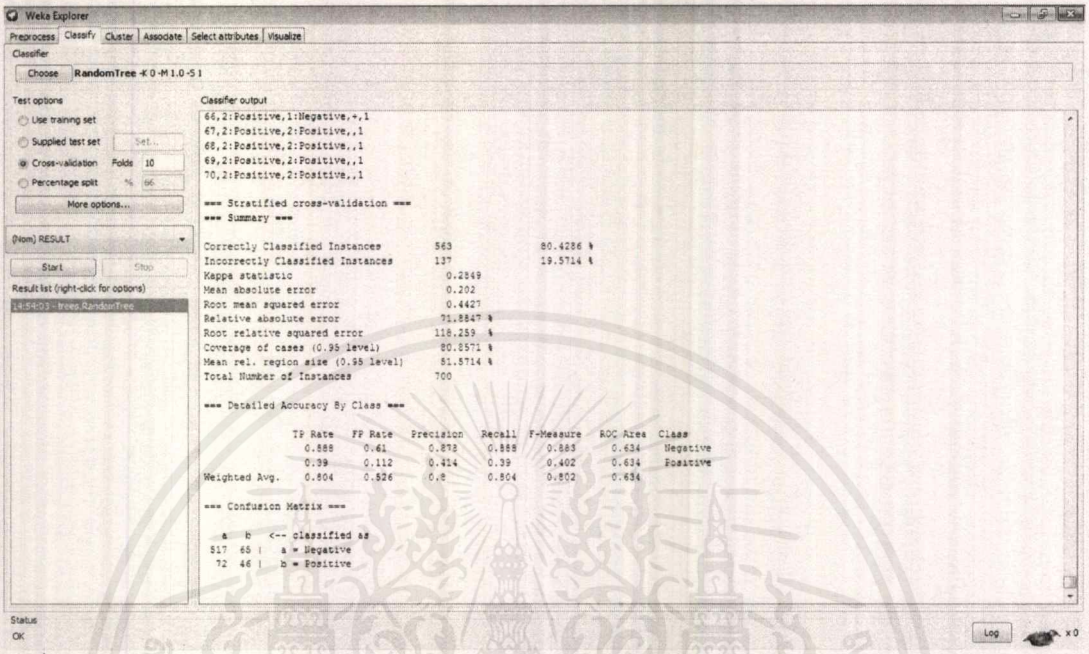


รูปที่ ข-41 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม Random Forest

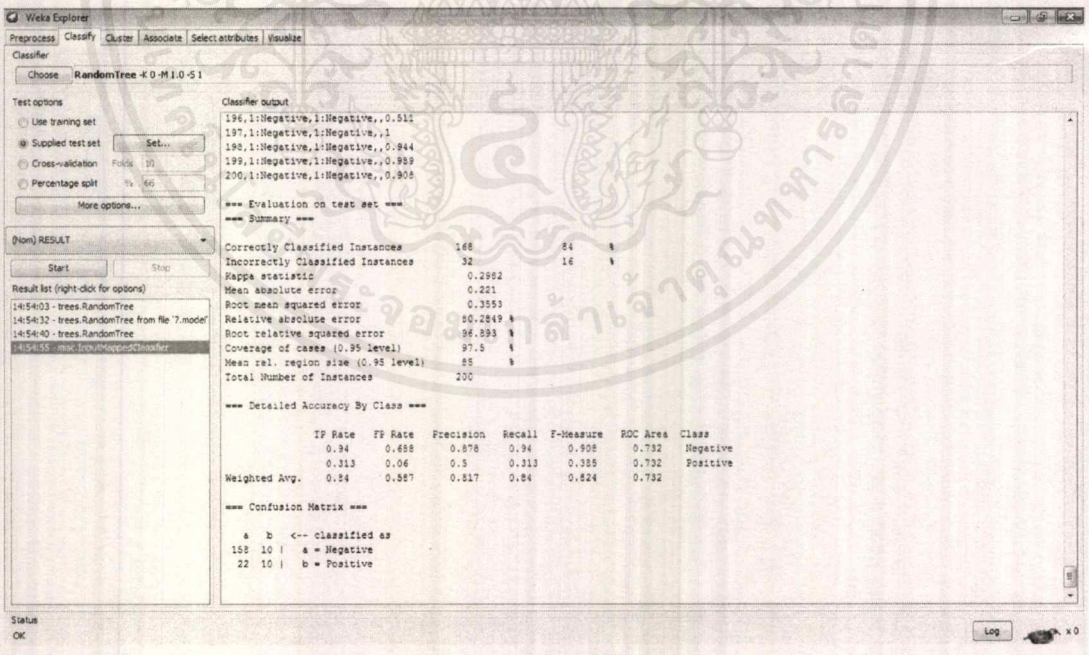


รูปที่ ข-42 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม Random Forest

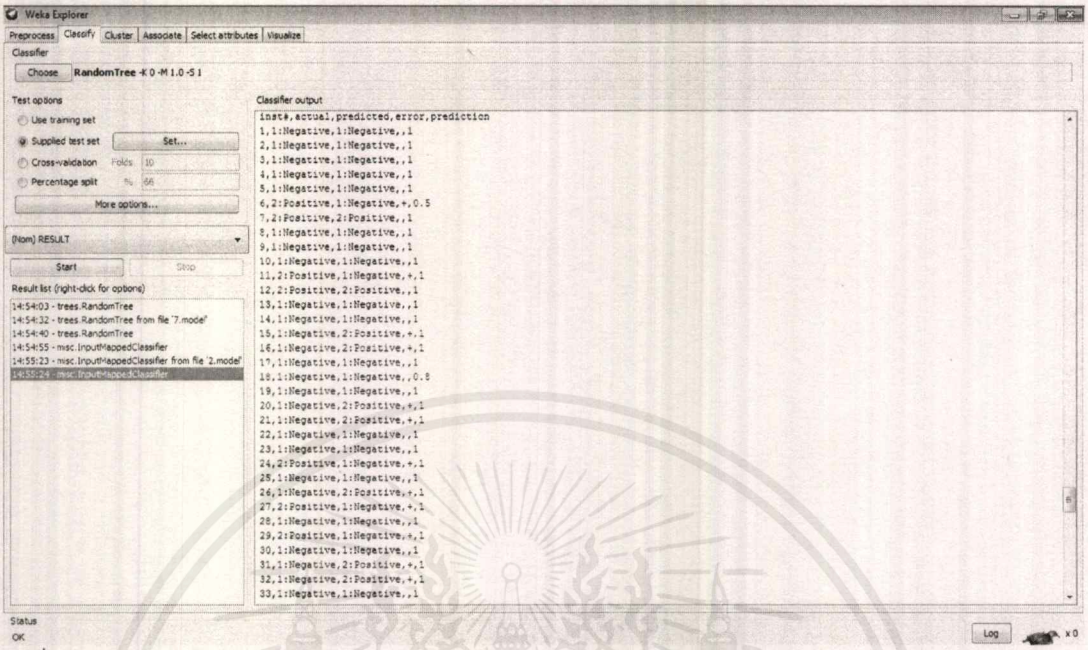
2.5 อัลกอริทึม Random Tree



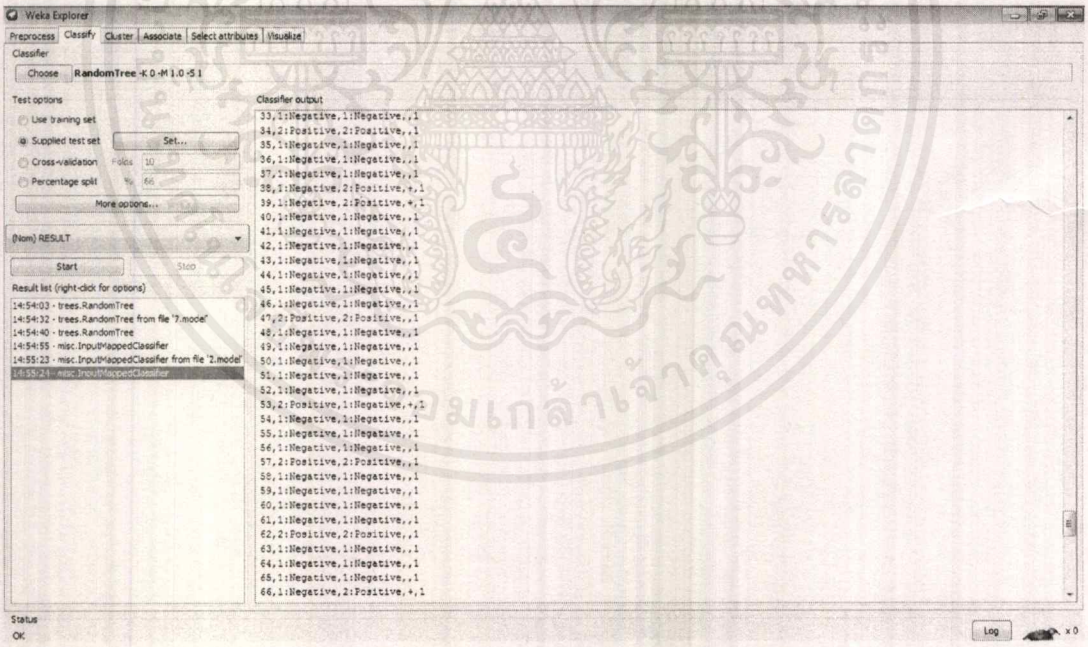
รูปที่ ข-43 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการสร้างตัวแบบโดยอัลกอริทึม Random Forest



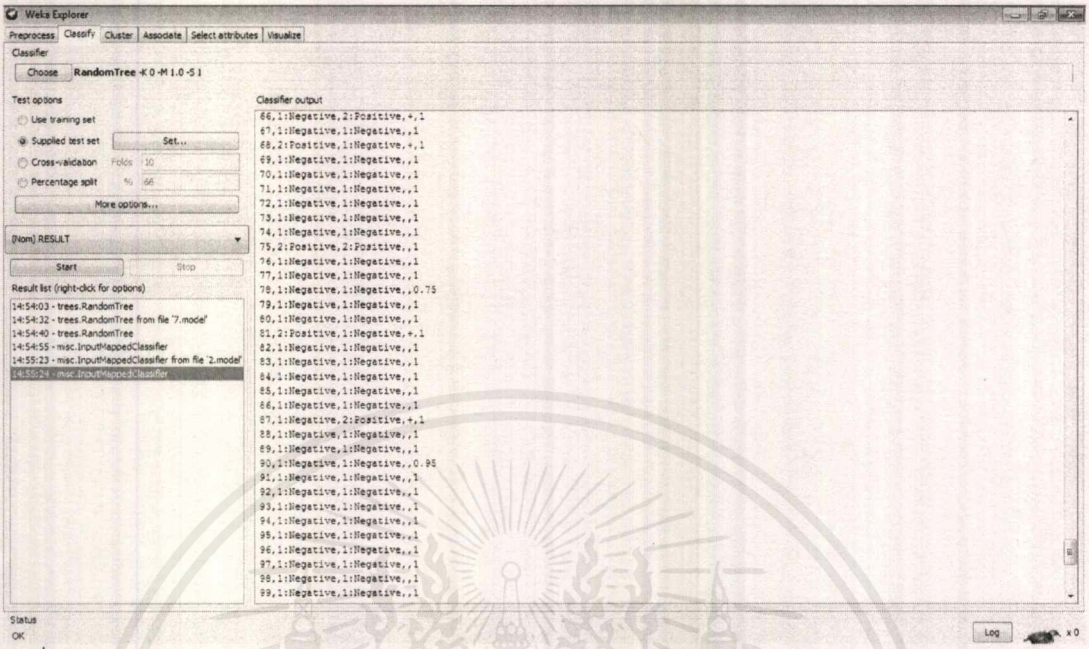
รูปที่ ข-44 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทดสอบตัวแบบโดยอัลกอริทึม Random Tree



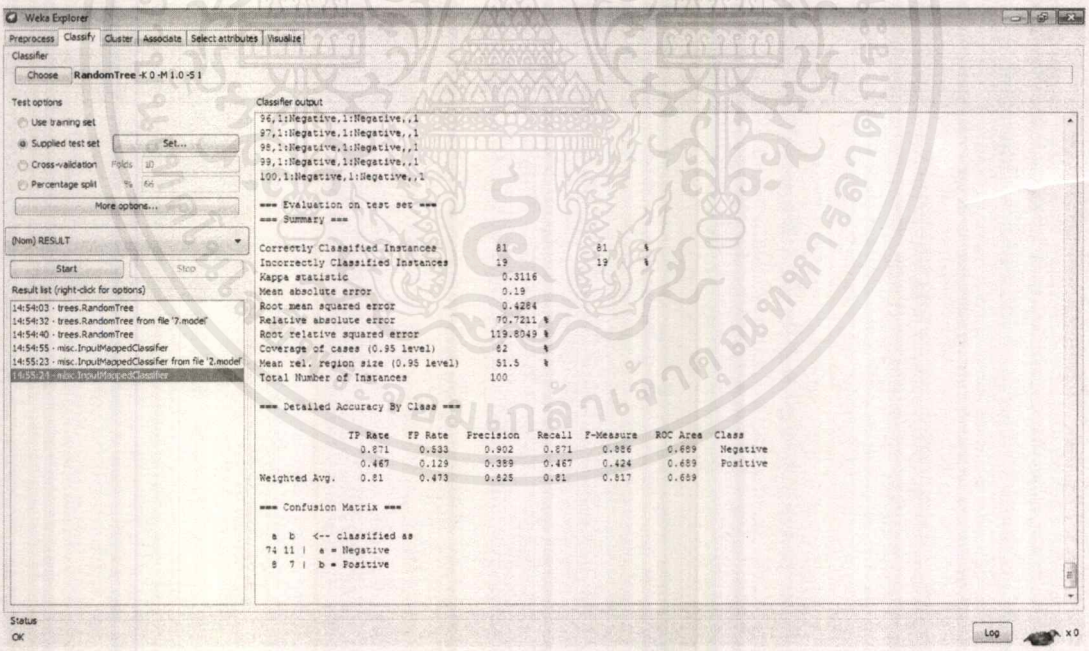
รูปที่ ข-45 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม Random Tree



รูปที่ ข-46 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม Random Tree

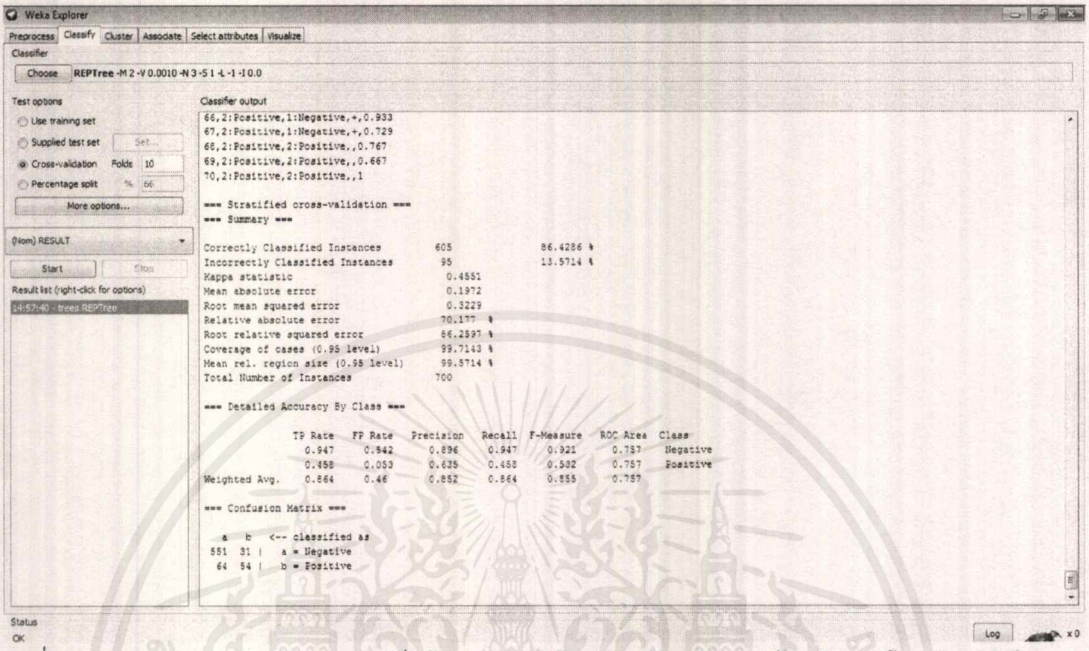


รูปที่ ข-47 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม Random Tree

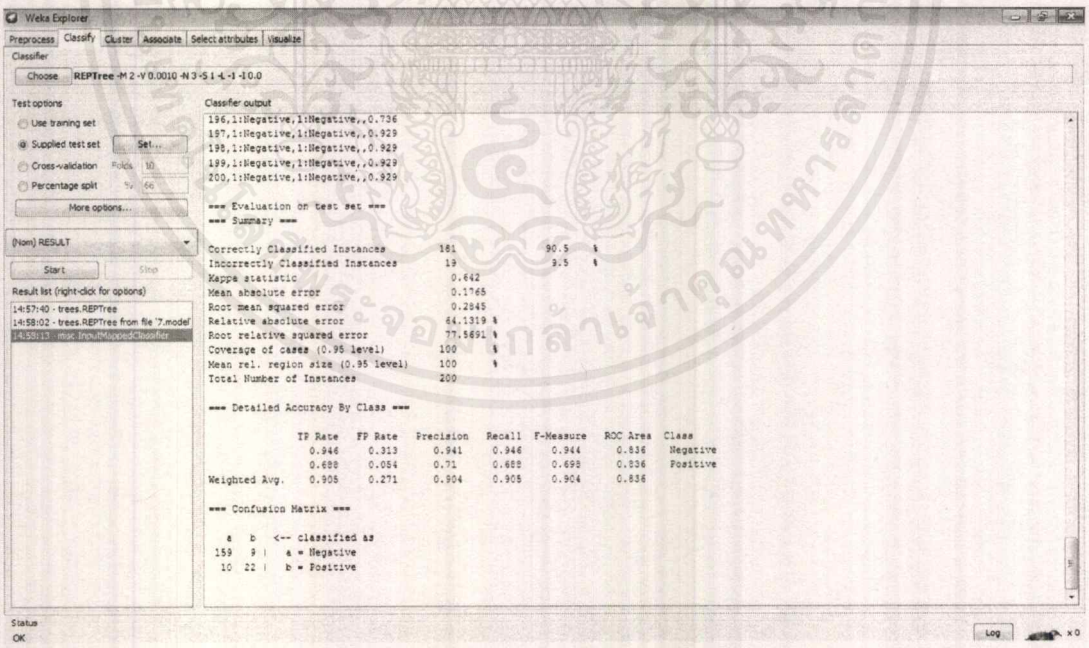


รูปที่ ข-48 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม Random Tree

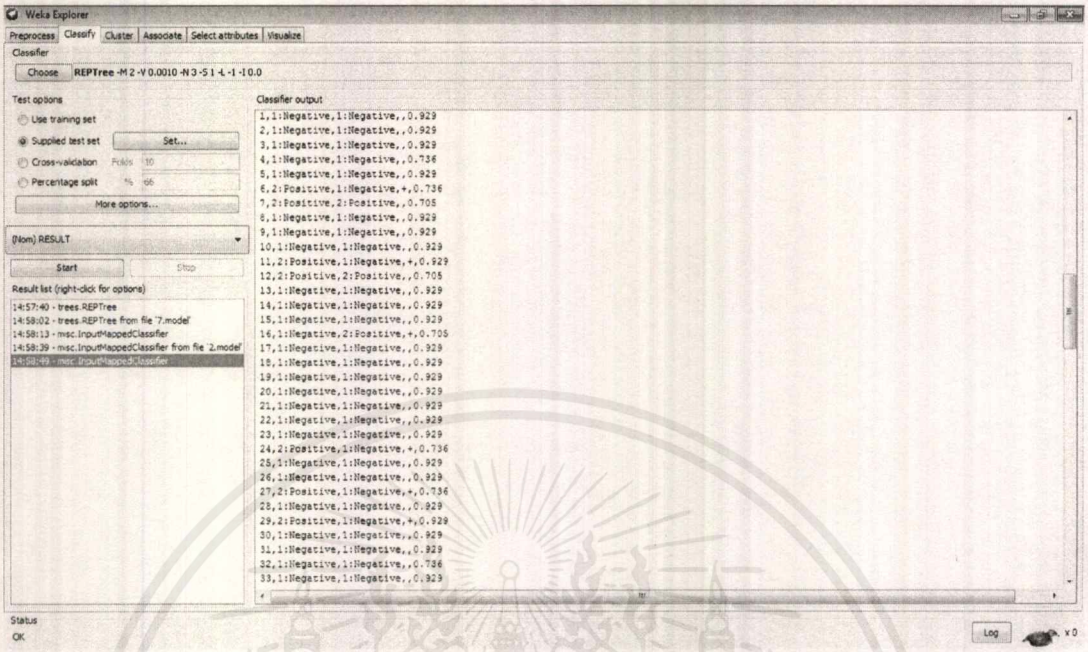
2.6 อัลกอริทึม REP Tree



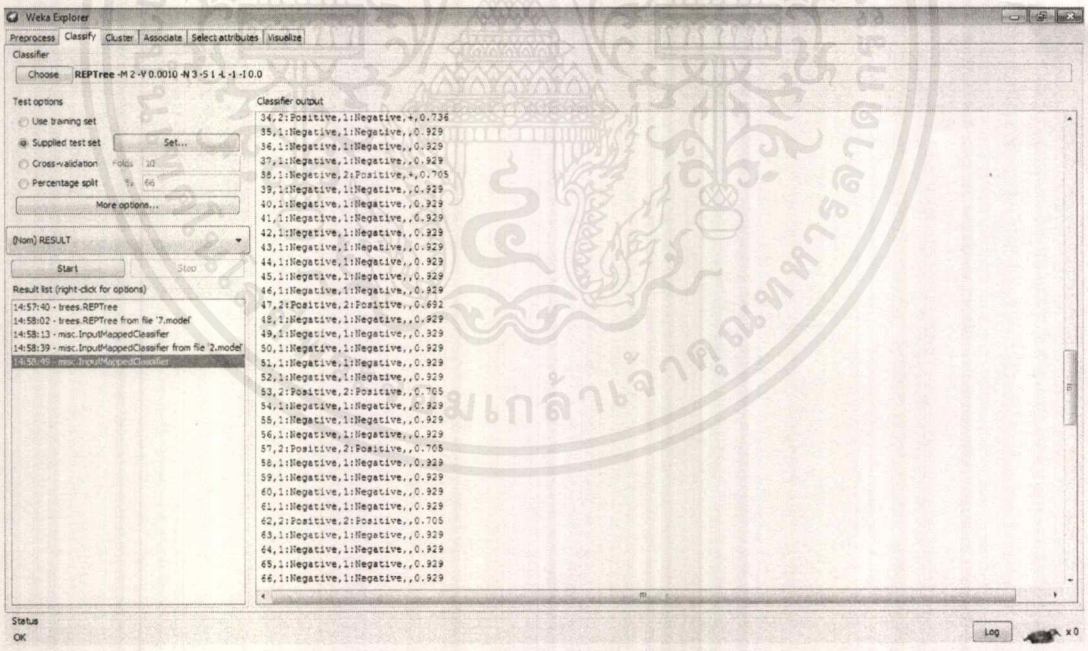
รูปที่ ข-49 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการสร้างตัวแบบโดยอัลกอริทึม REP Tree



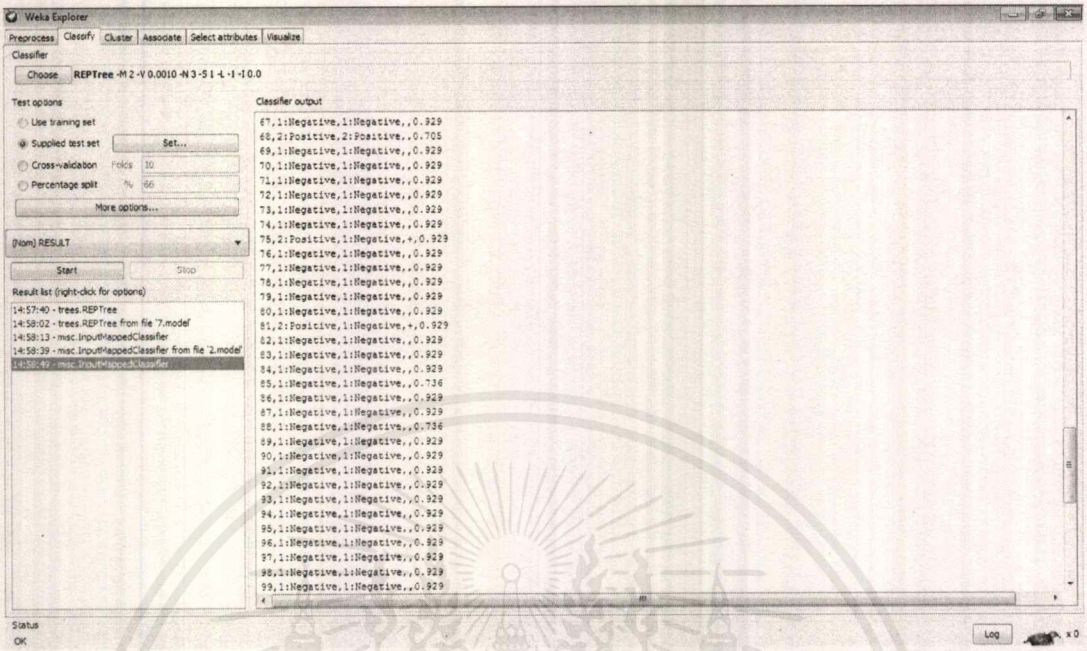
รูปที่ ข-50 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทดสอบตัวแบบโดยอัลกอริทึม REP Tree



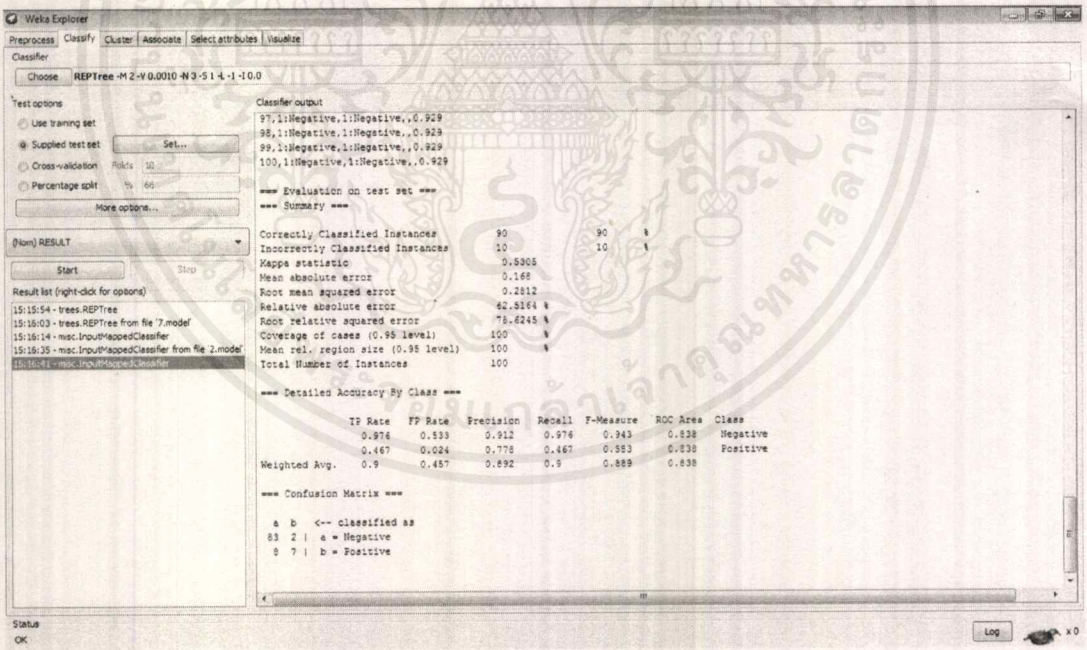
รูปที่ ข-51 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม REP Tree



รูปที่ ข-52 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม REP Tree

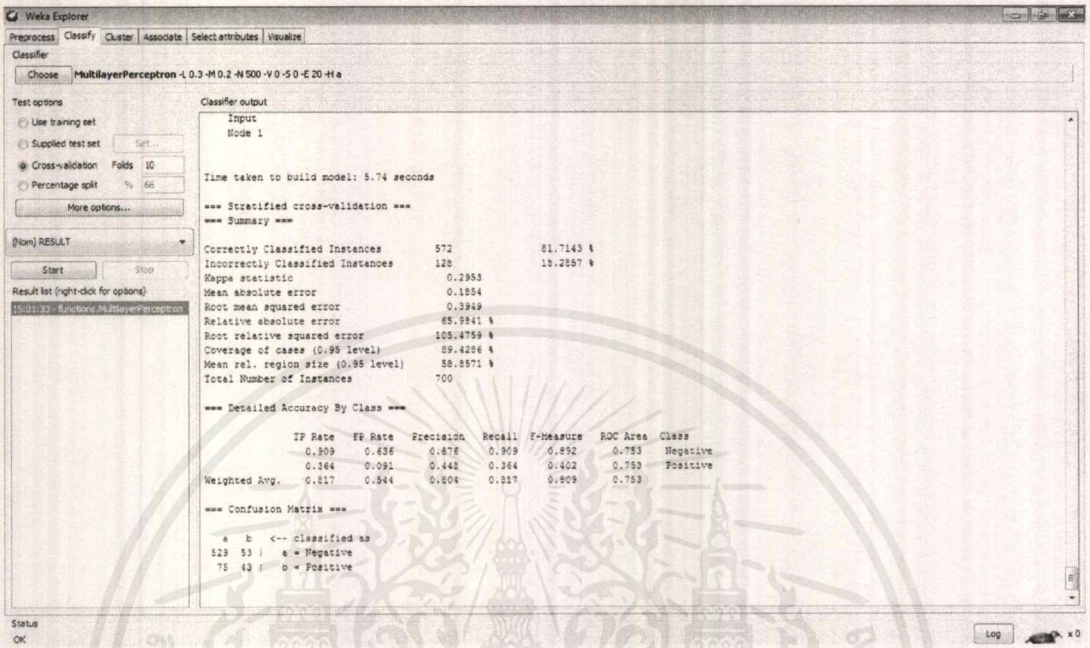


รูปที่ ข-53 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม REP Tree

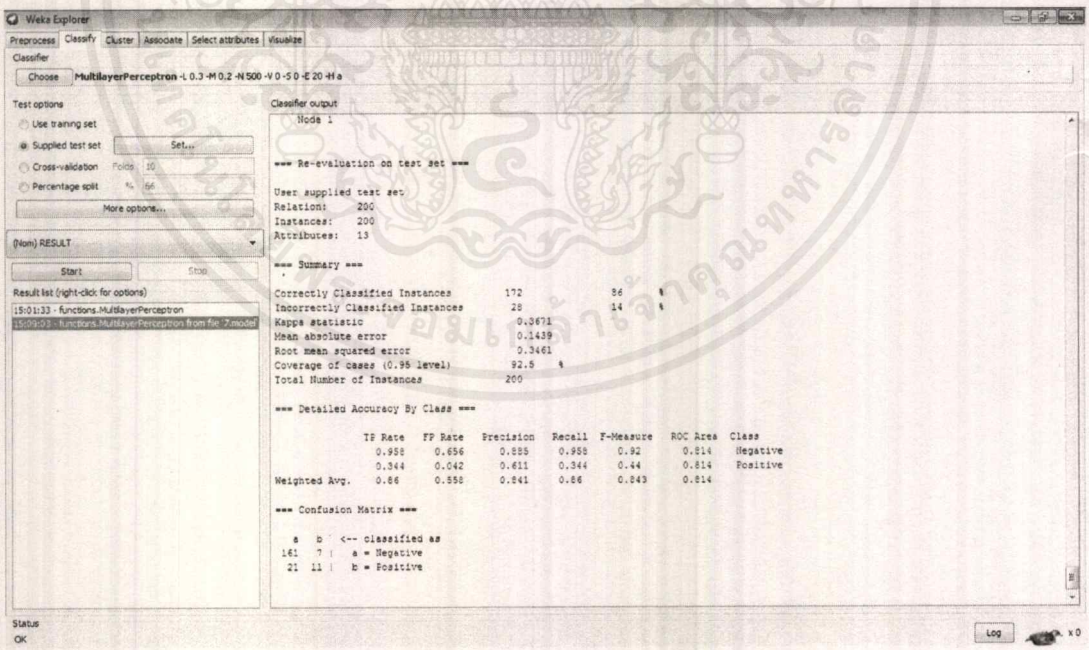


รูปที่ ข-54 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม REP Tree

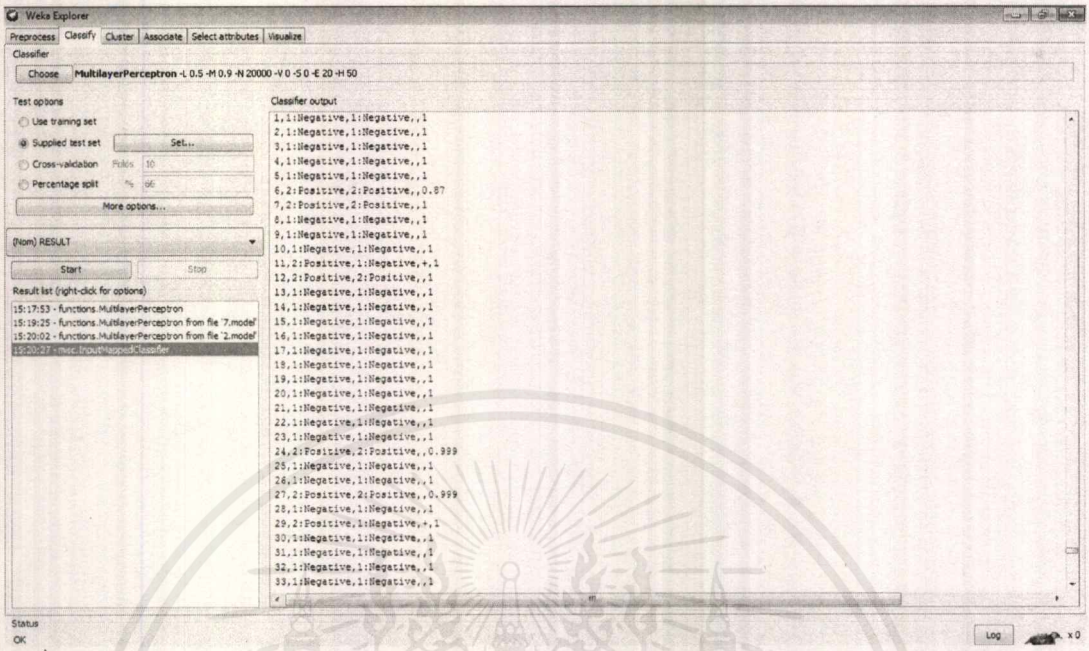
3. วิธีโครงข่ายประสาทเทียม



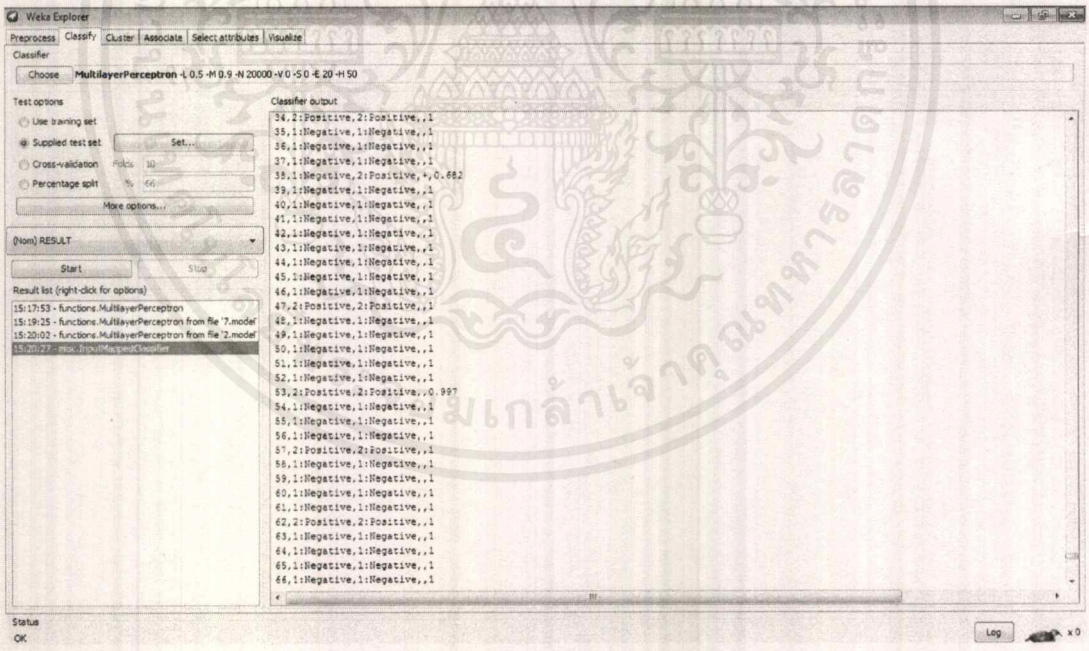
รูปที่ ข-55 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการสร้างตัวแบบโดยวิธีโครงข่ายประสาทเทียม



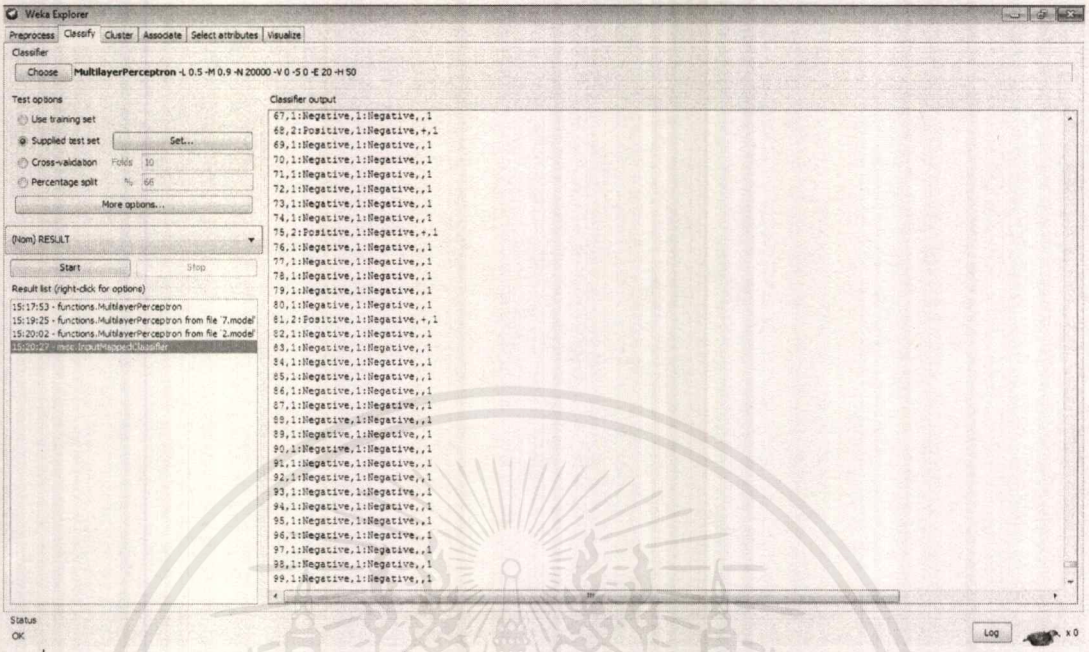
รูปที่ ข-56 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทดสอบตัวแบบโดยวิธีโครงข่ายประสาทเทียม



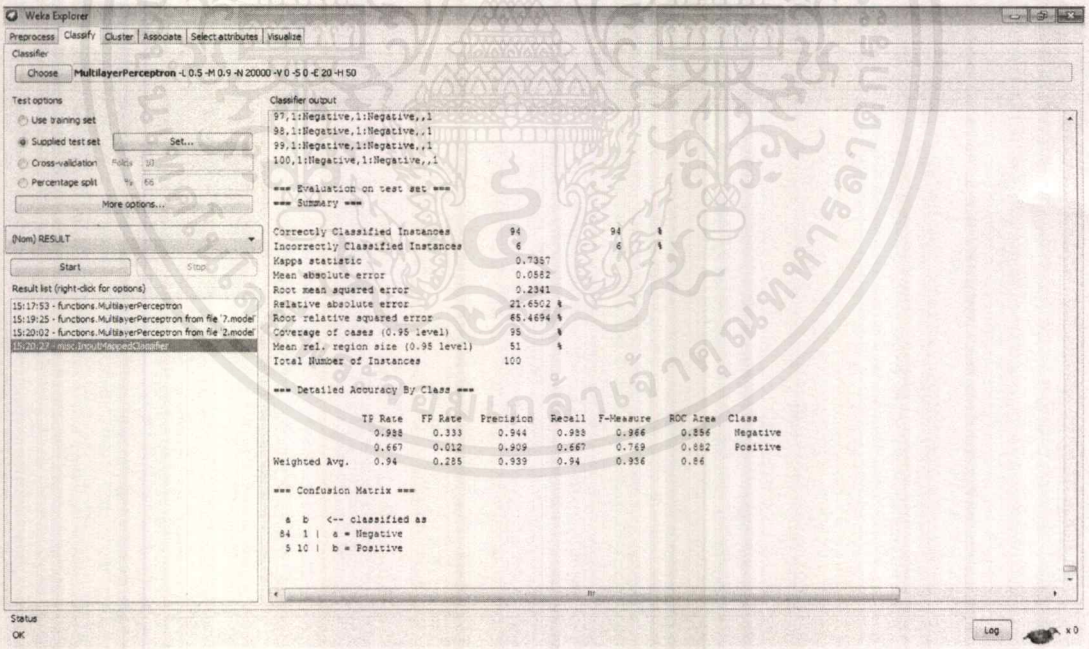
รูปที่ ข-57 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยวิธีโครงข่ายประสาทเทียม



รูปที่ ข-58 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยวิธีโครงข่ายประสาทเทียม



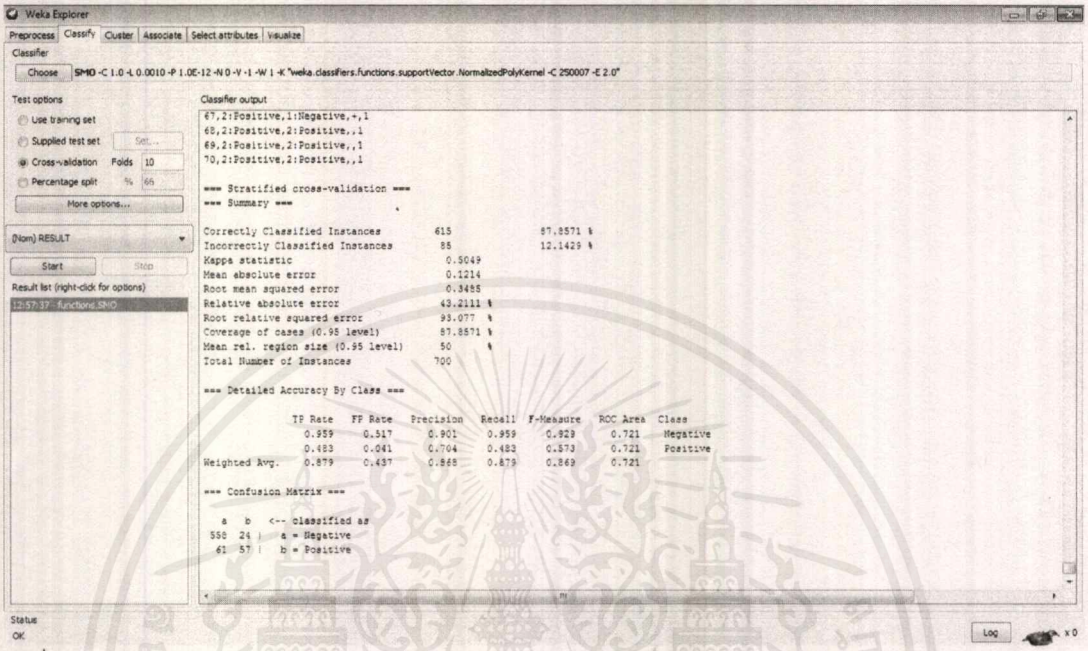
รูปที่ ข-59 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยวิธีโครงข่ายประสาทเทียม



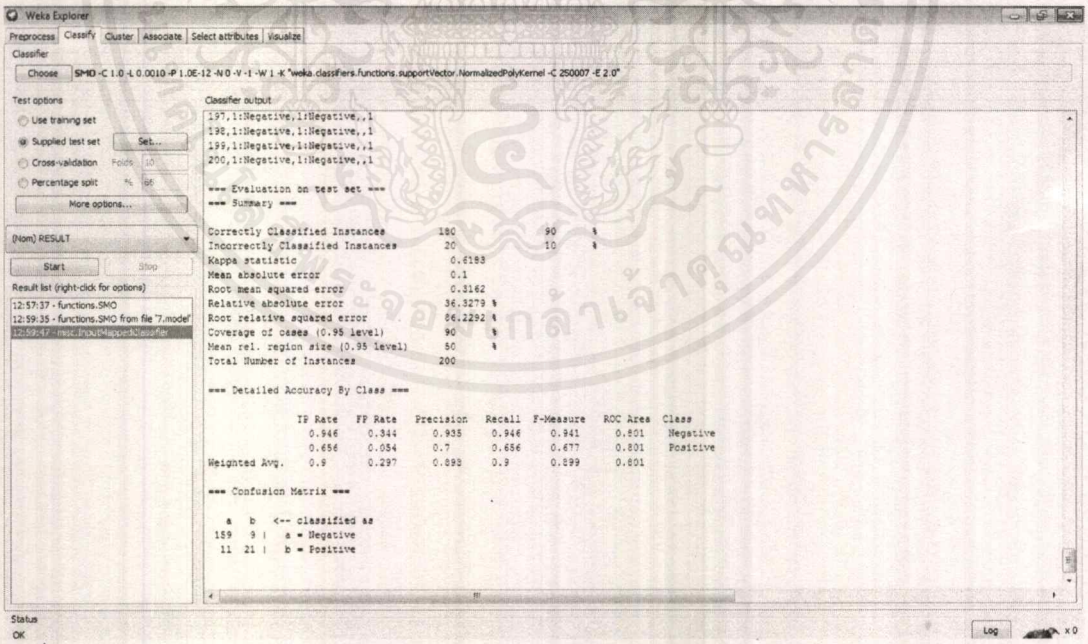
รูปที่ ข-60 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยวิธีโครงข่ายประสาทเทียม

4. วิธีซัพพอร์ตเวกเตอร์แมชชีน

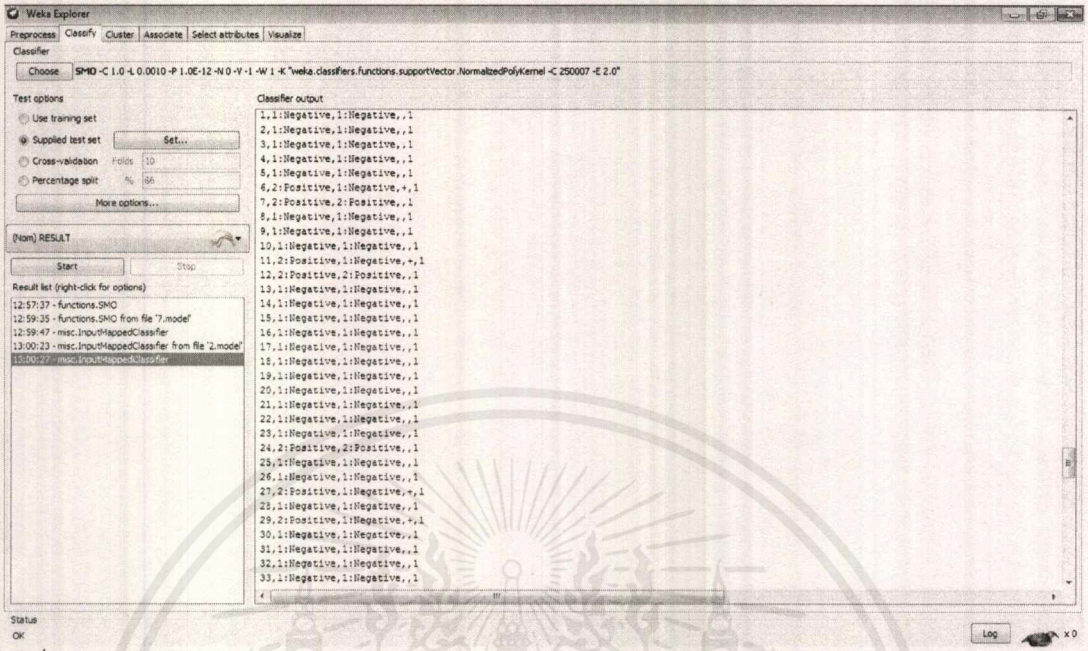
4.1 อัลกอริทึม Normalized Poly Kernel



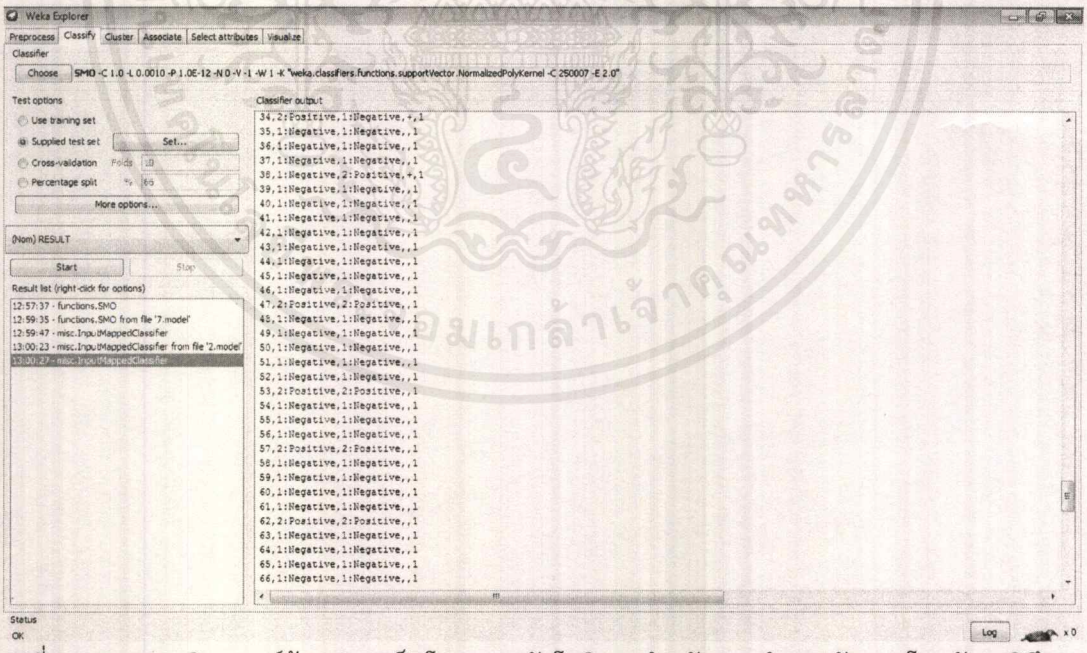
รูปที่ ข-61 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการสร้างตัวแบบโดยอัลกอริทึม Normalized Poly Kernel



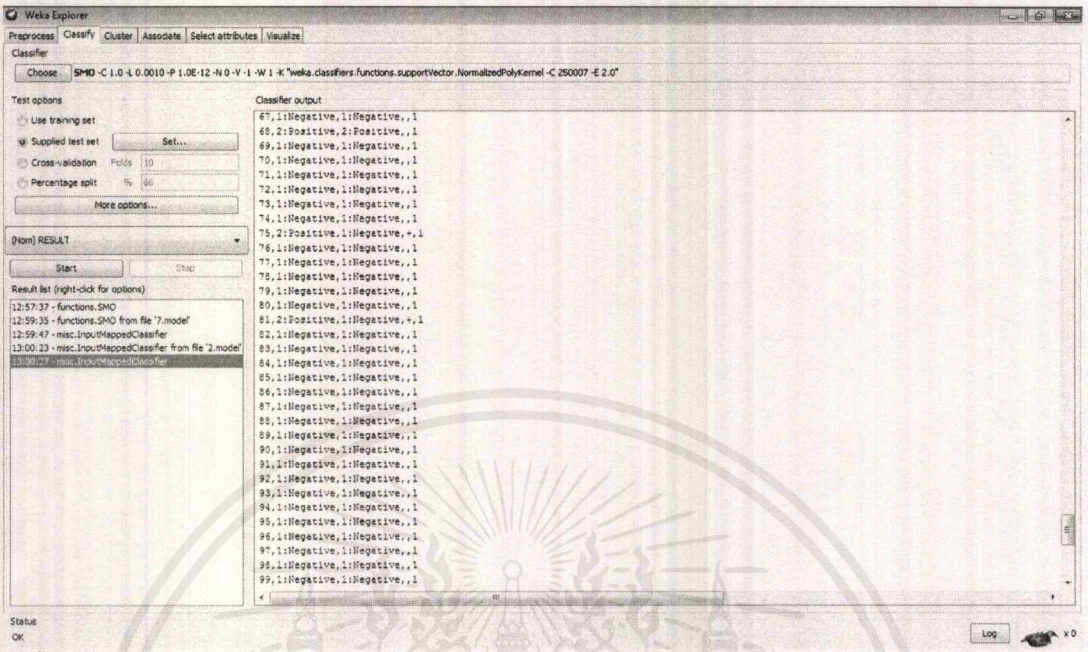
รูปที่ ข-62 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทดสอบตัวแบบโดยอัลกอริทึม Normalized Poly Kernel



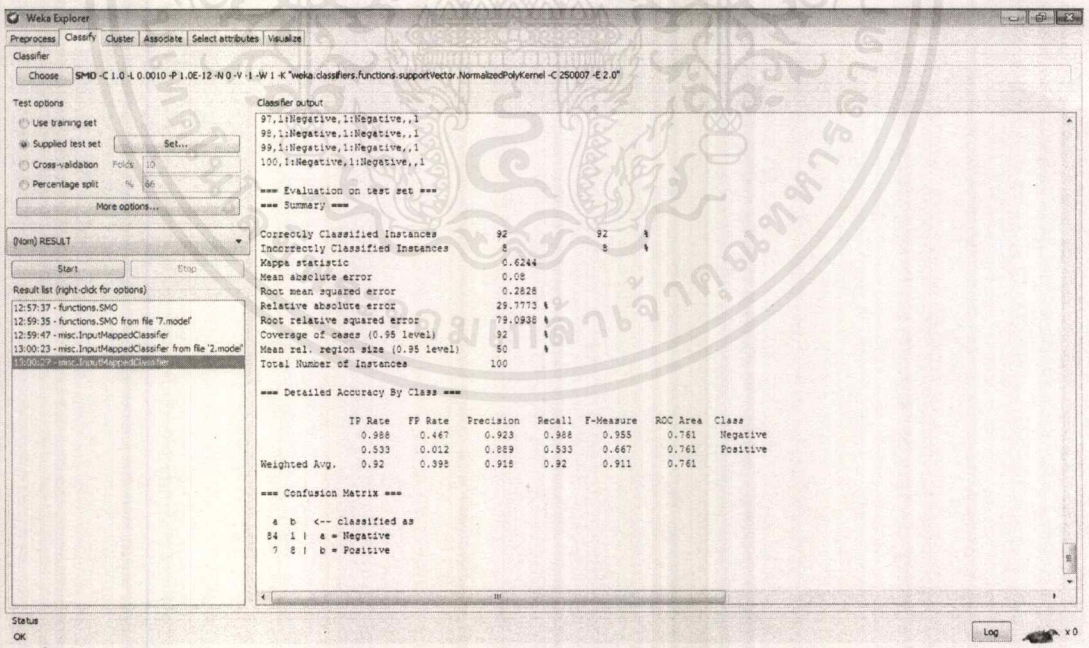
รูปที่ ข-63 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม Normalized Poly Kernel



รูปที่ ข-64 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม Normalized Poly Kernel

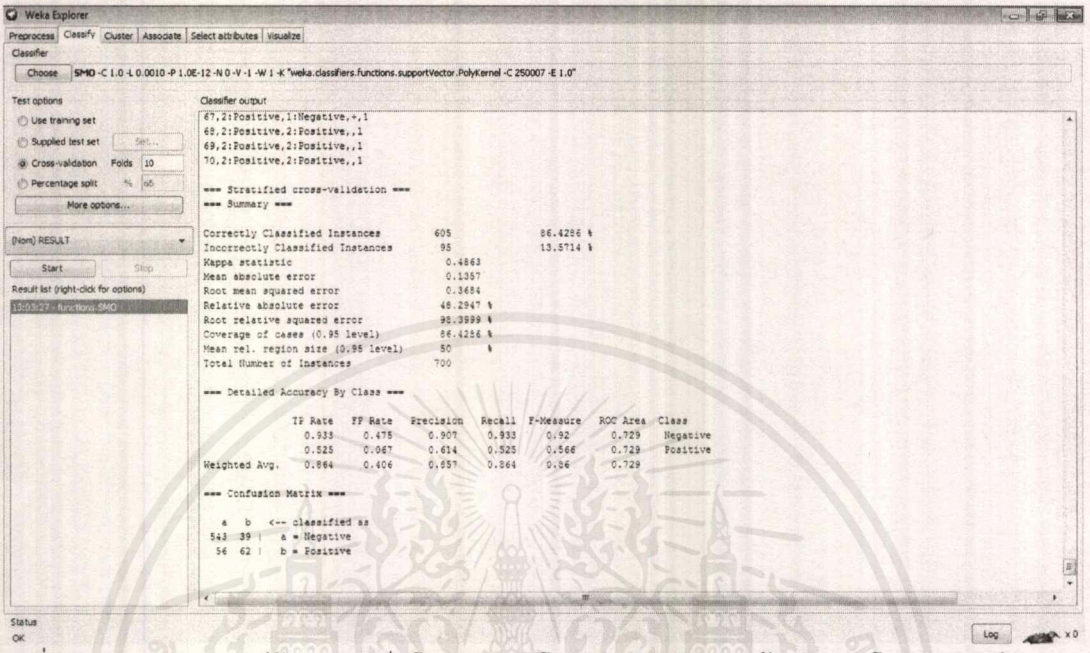


รูปที่ ข-65 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม Normalized Poly Kernel

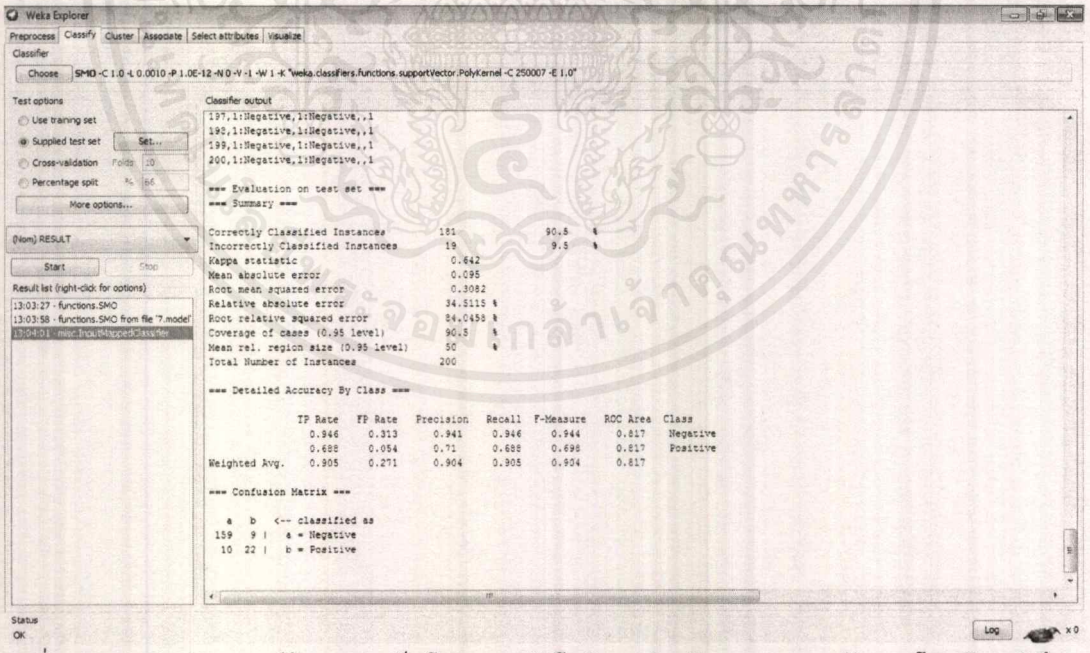


รูปที่ ข-66 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม Normalized Poly Kernel

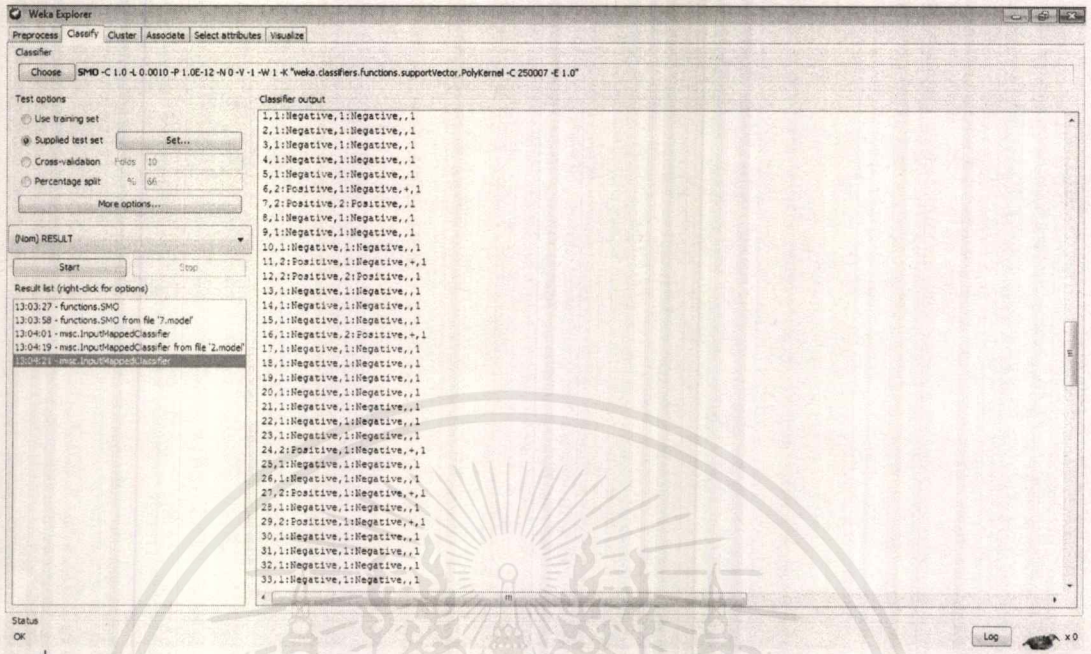
4.2 อัลกอริทึม Poly Kernel



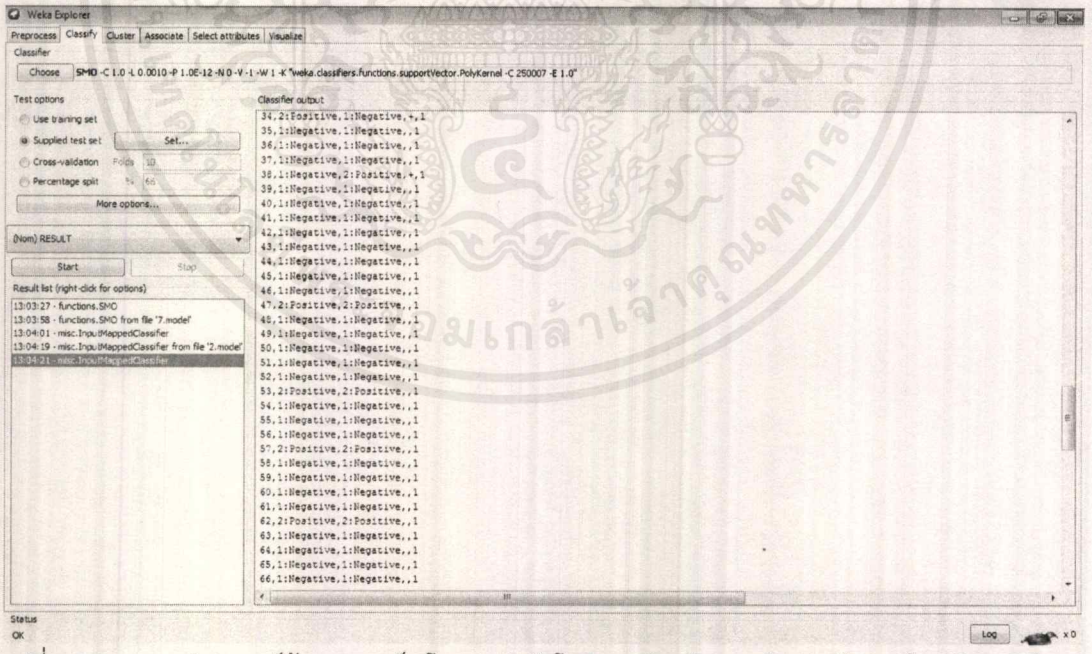
รูปที่ ข-6 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการสร้างตัวแบบโดยอัลกอริทึม Poly Kernel



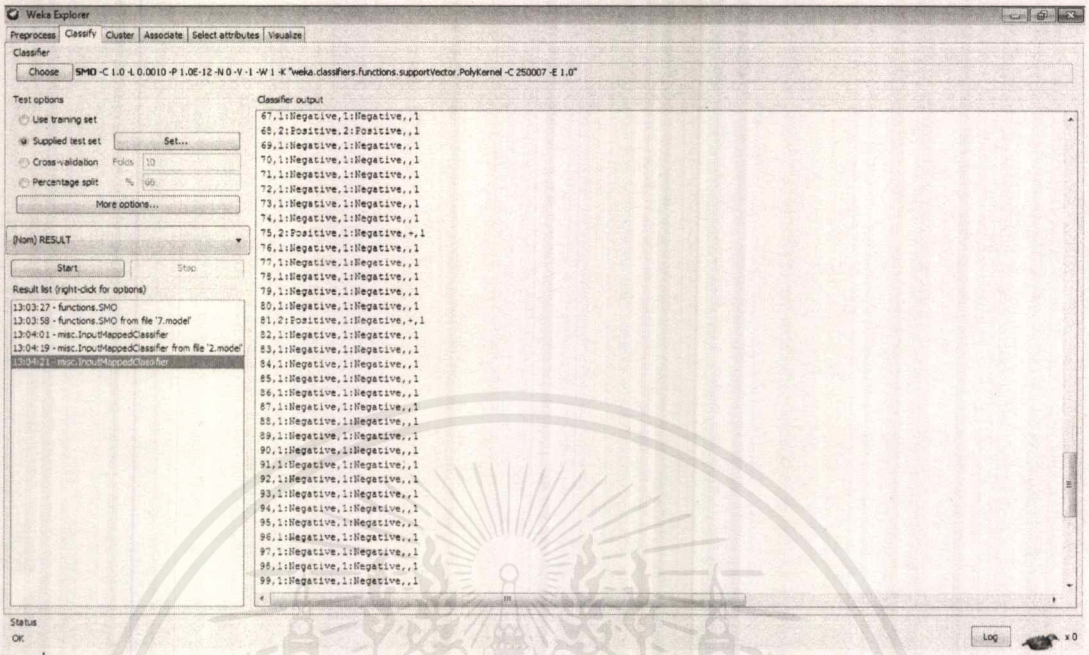
รูปที่ ข-68 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทดสอบตัวแบบโดยอัลกอริทึม Poly Kernel



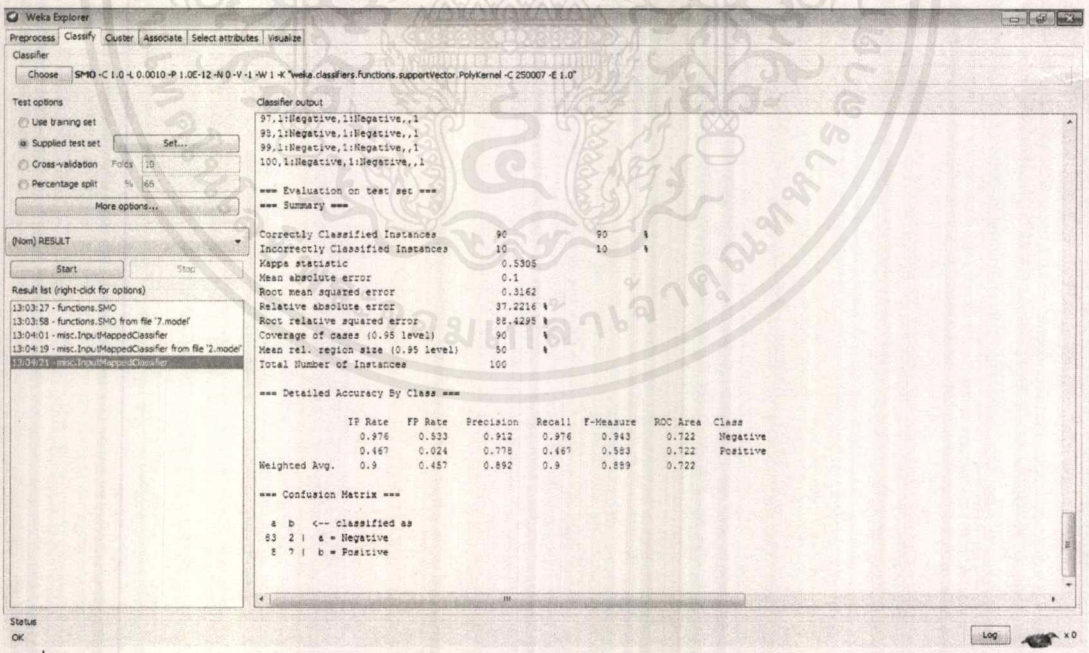
รูปที่ ข-69 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม Poly Kernel



รูปที่ ข-70 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม Poly Kernel

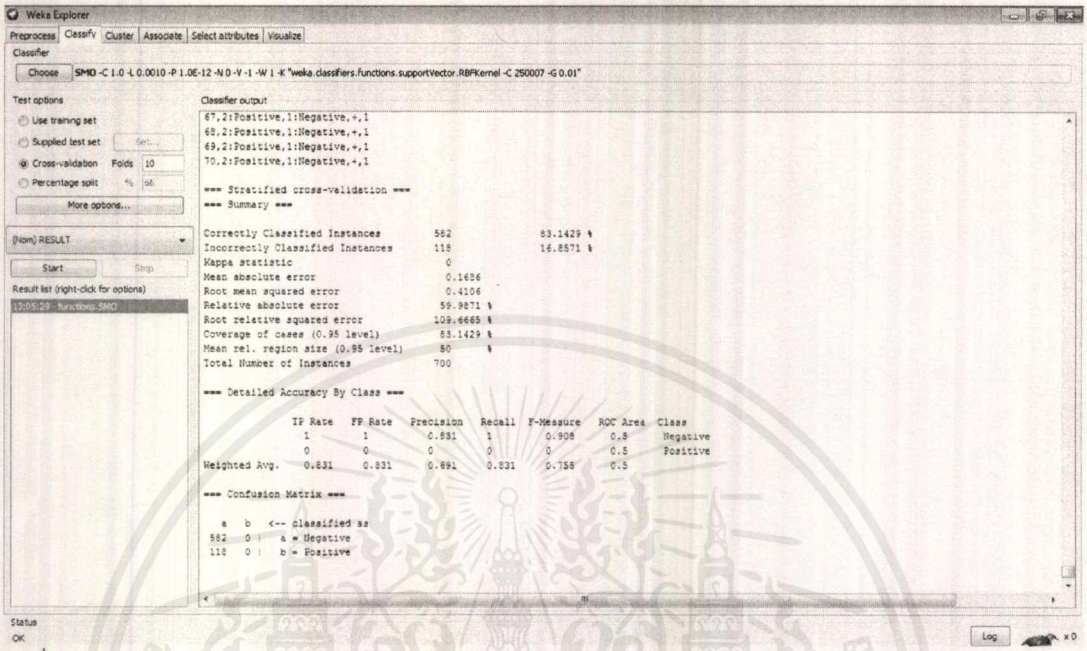


รูปที่ ข-71 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม Poly Kernel

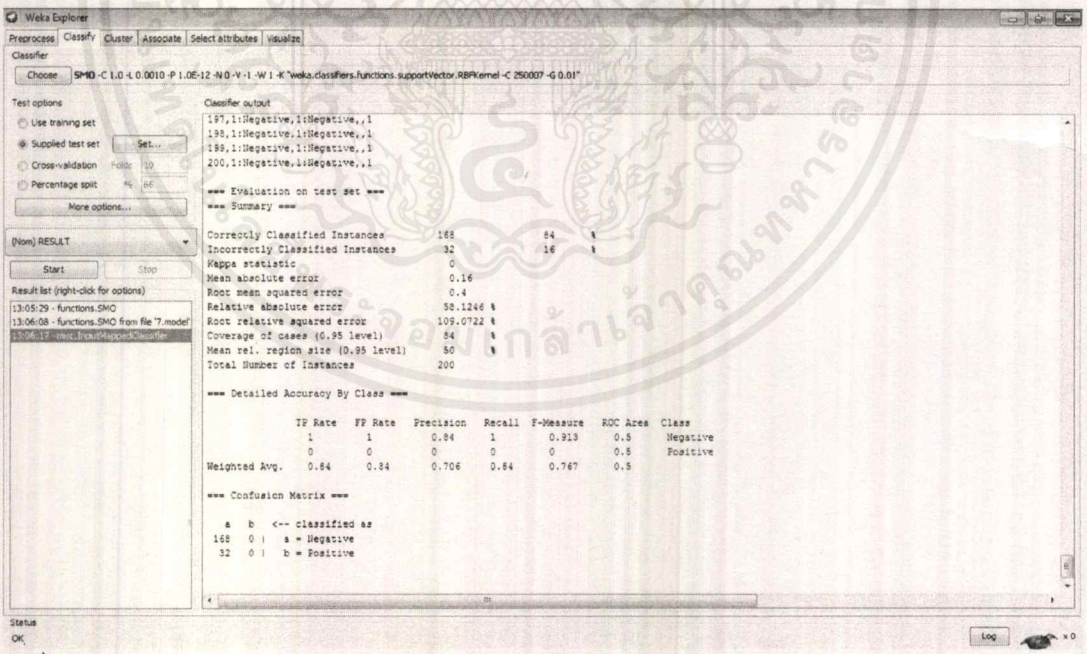


รูปที่ ข-72 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม Poly Kernel

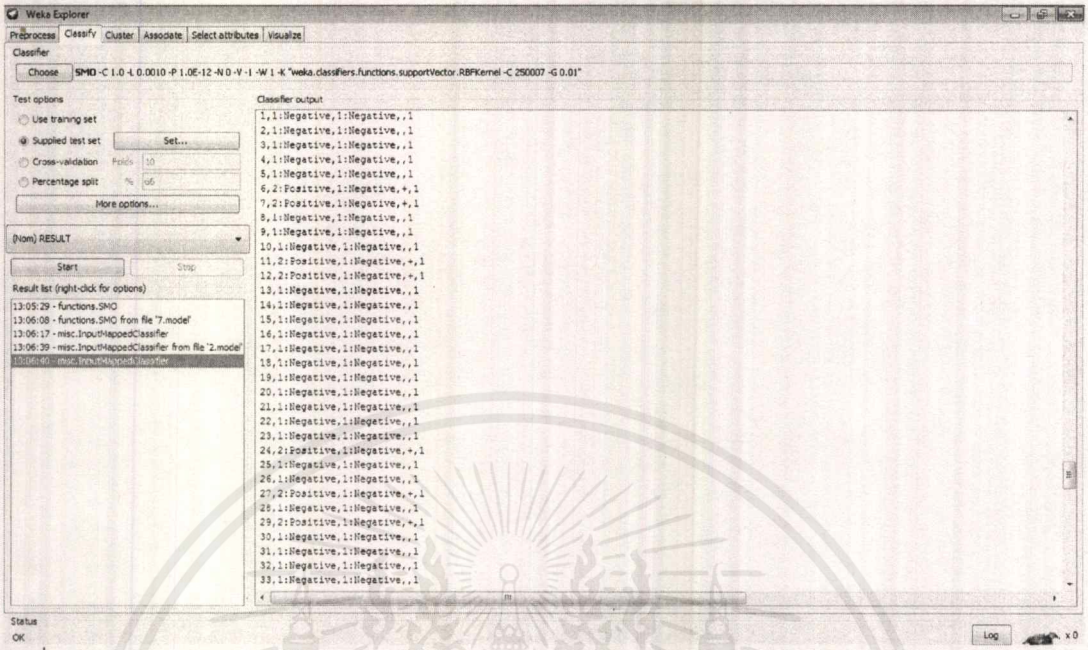
4.3 อัลกอริทึม RBF Kernel



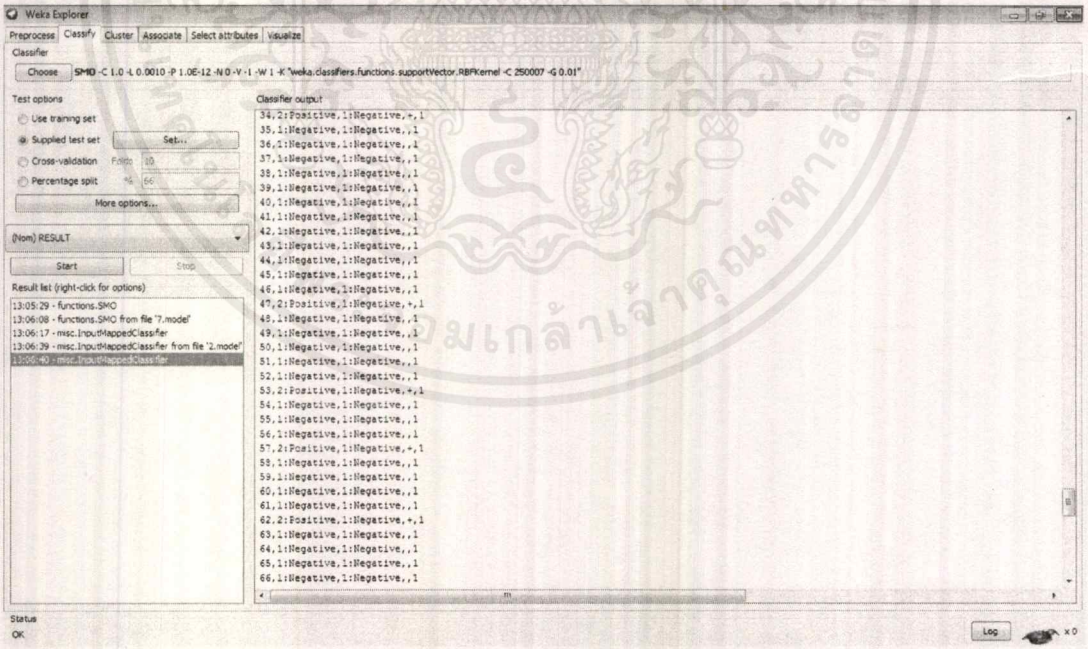
รูปที่ ข-73 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการสร้างตัวแบบโดยอัลกอริทึม RBF Kernel



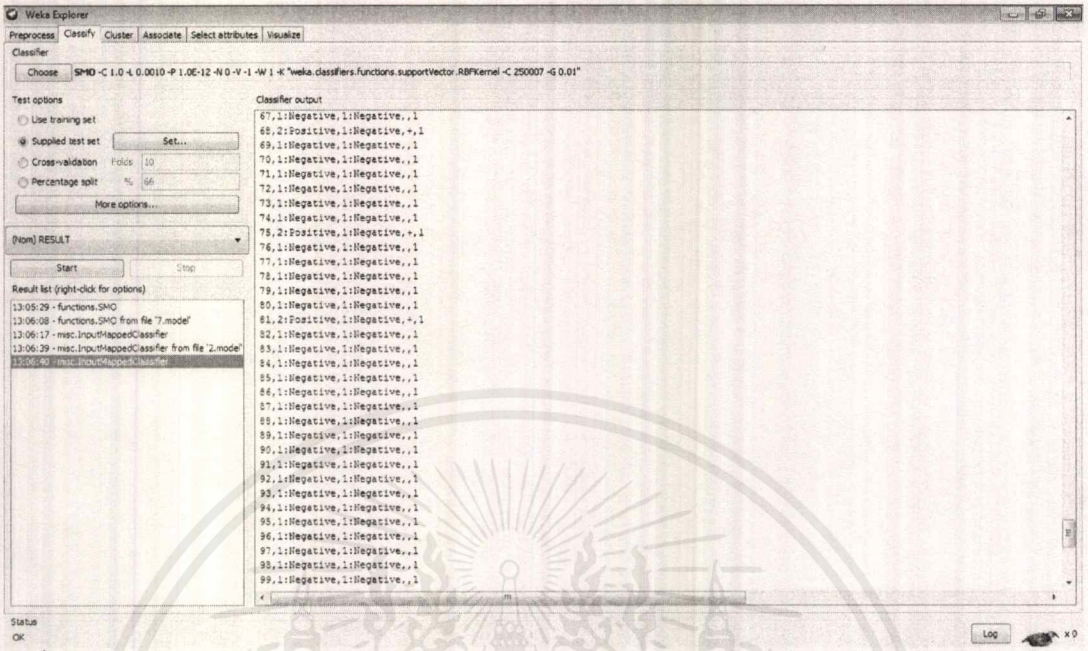
รูปที่ ข-74 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทดสอบตัวแบบโดยอัลกอริทึม RBF Kernel



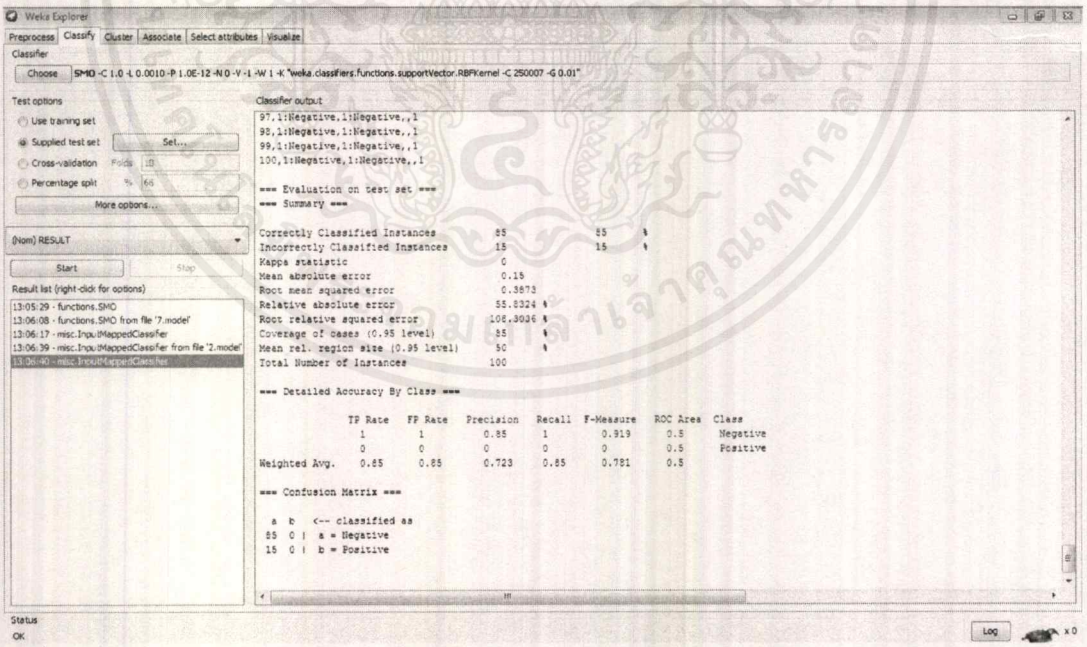
รูปที่ ข-75 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม RBF Kernel



รูปที่ ข-76 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม RBF Kernel

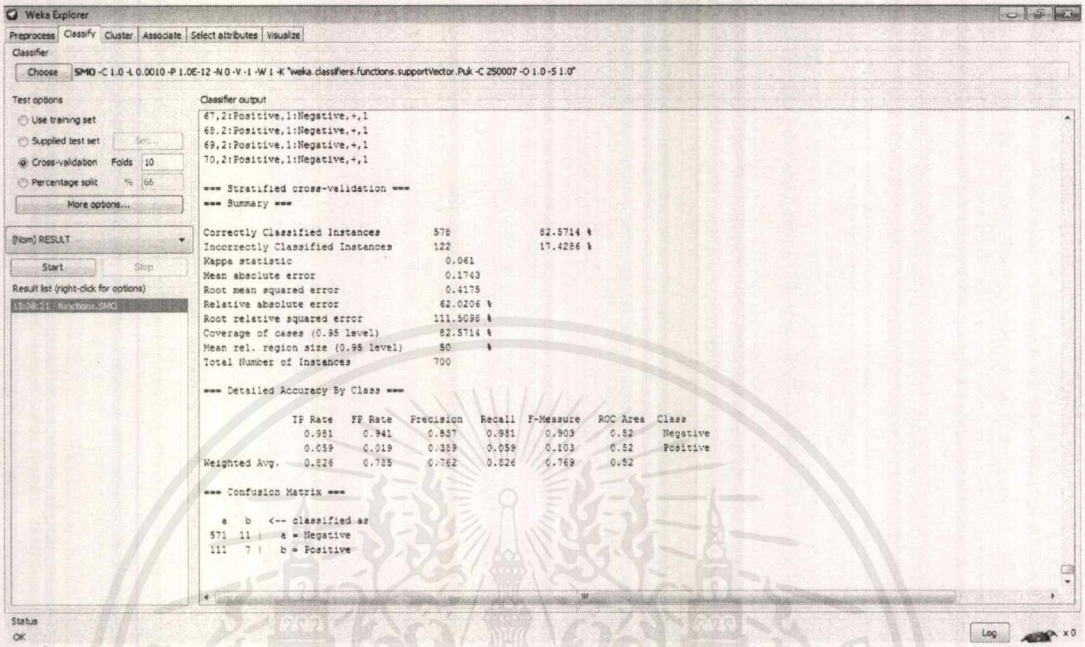


รูปที่ ข-77 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม RBF Kernel

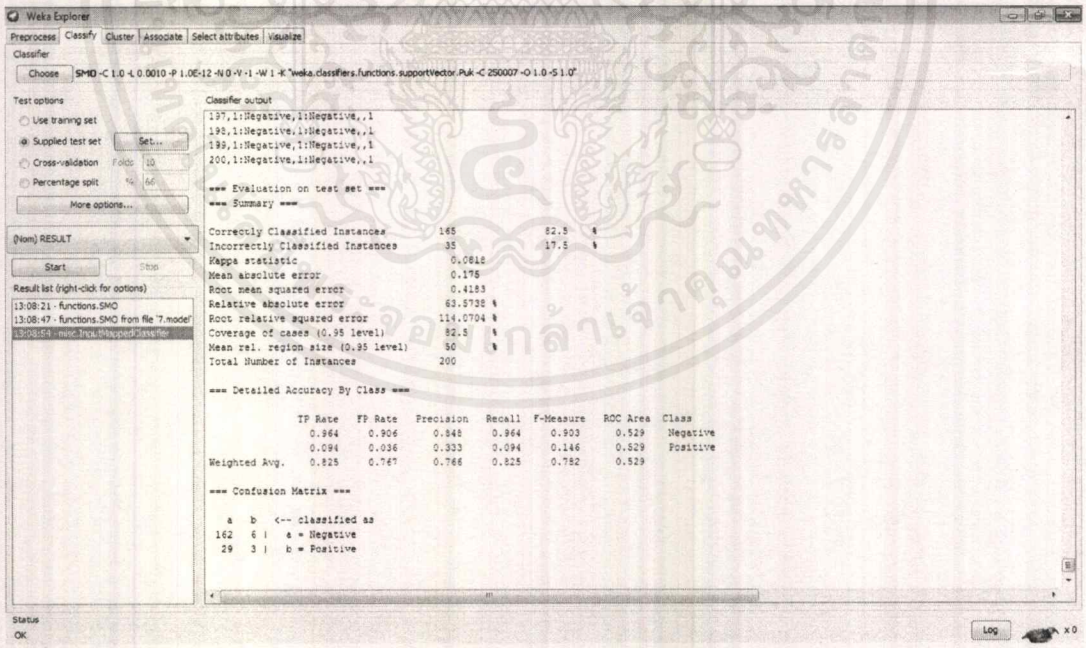


รูปที่ ข-78 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม RBF Kernel

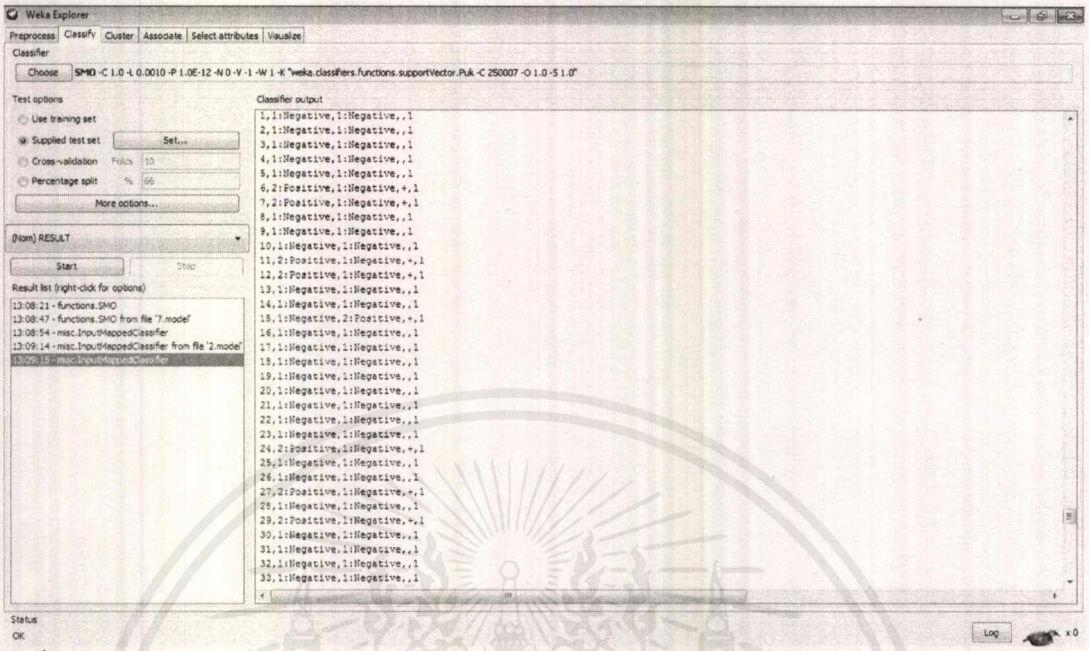
4.4 อัลกอริทึม Puk



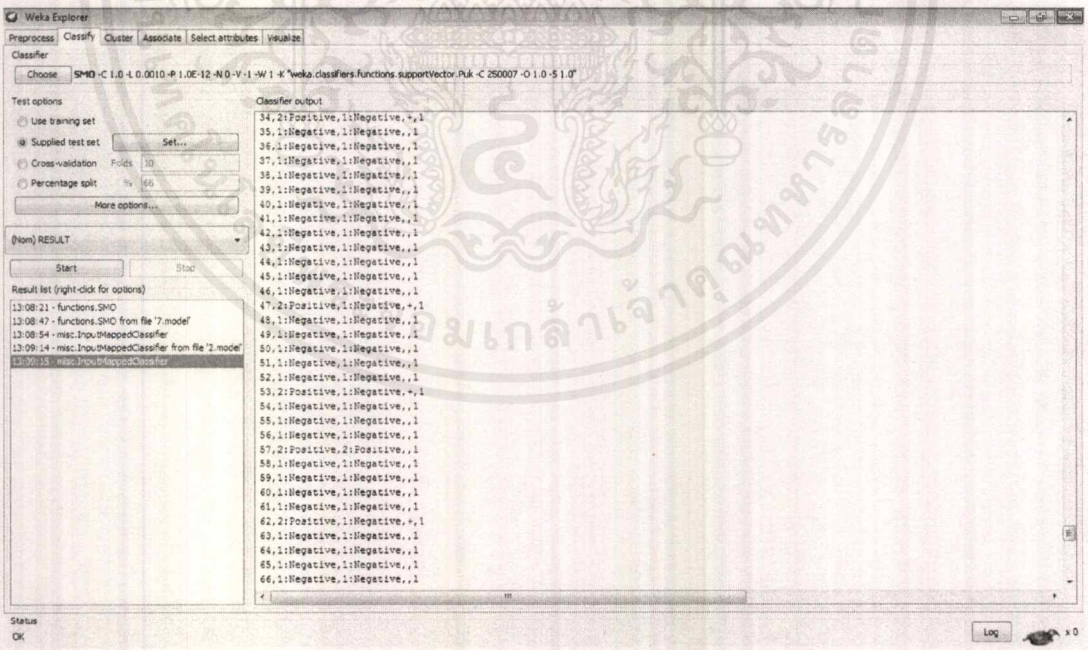
รูปที่ ข-79 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการสร้างตัวแบบโดยอัลกอริทึม Puk



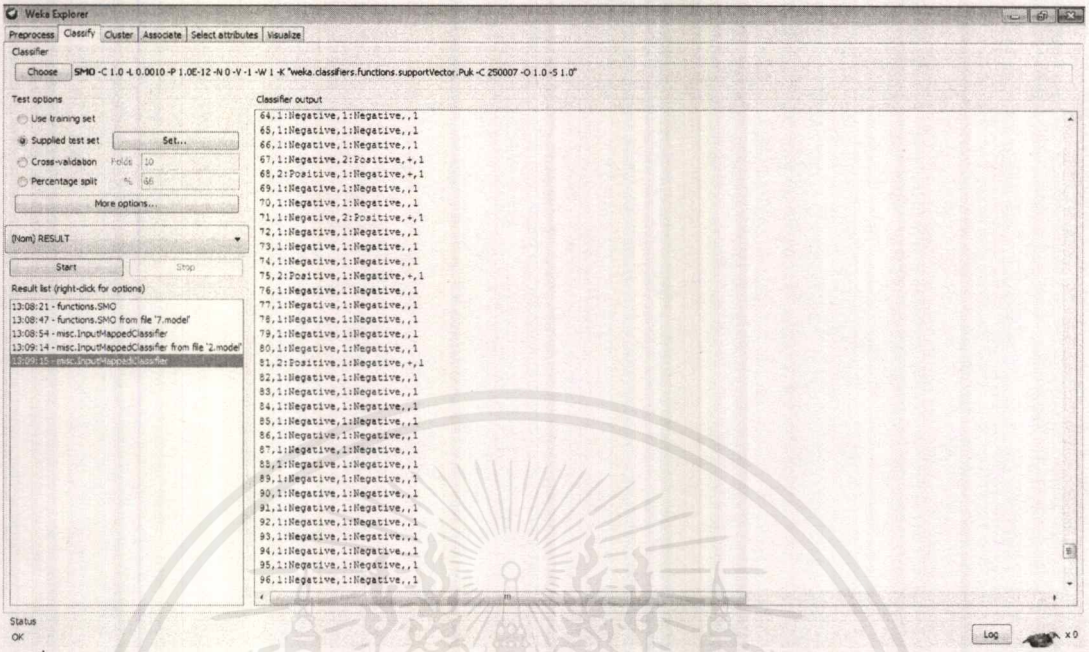
รูปที่ ข-80 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทดสอบตัวแบบโดยอัลกอริทึม Puk



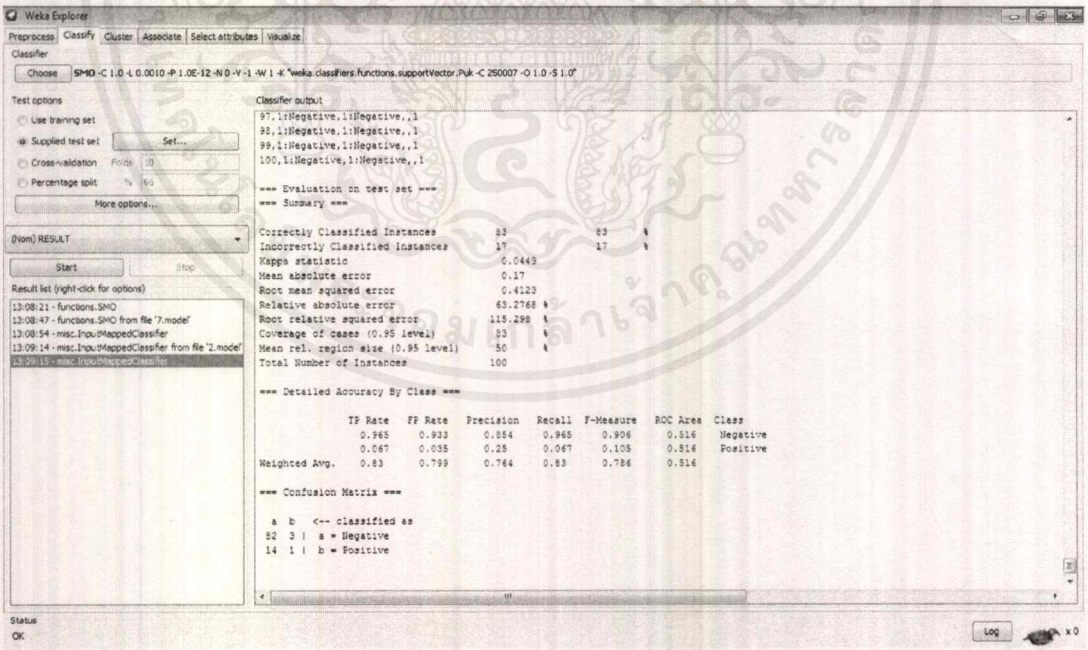
รูปที่ ข-81 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม Puk



รูปที่ ข-82 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม Puk



รูปที่ ข-83 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม Puk



รูปที่ ข-84 ผลการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยอัลกอริทึม Puk

ภาคผนวก ค

ตัวอย่างการคำนวณ

<u>ตัวอย่างที่ 1</u> การคำนวณค่าความถูกต้อง(Accuracy) ค่าความแม่นยำ (Precision)ค่าความระลึก (Recall)ค่าความถ่วงดุล (F-Measure)ของการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูง โดยวิธีความใกล้เคียงกันมากที่สุด

จากรูปที่ ข-6

ค่าความถูกต้อง(Accuracy) $= \frac{TP+TN}{TP+TN+FP+FN}$

$= \frac{72+7}{72+7+8+13}$

$=$ 0.79 หรือ 79%

ค่าความแม่นยำ (Precision) $= \frac{TP}{TP+FP}$

$= \frac{72}{72+8}$

$=$ 0.9หรือ 90%

ค่าความระลึก (Recall) $= \frac{TP}{TP+FN}$

$= \frac{72}{72+13}$

$=$ 0.8471หรือ 84.71%

ค่าความถ่วงดุล(F-Measure) $= \frac{2\times(Recall\times Precision)}{Recall+Precision}$

$= \frac{2\times(0.8471\times 0.9)}{0.8471+0.9}$

$=$ 0.8727หรือ 87.27%

<u>ตัวอย่างที่ 2</u> การคำนวณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) ของการวิเคราะห์ข้อมูลการเป็นโรคความดันโลหิตสูงสำหรับการทำนายตัวแบบโดยวิธีความใกล้เคียงกันมากที่สุด

จากรูปที่ ข-3 ถึง ข-6

กำหนดให้ค่า $y_i = 1$ ได้จากกรณีที่ ค่าจริง (actual) ใน class attribute ของข้อมูลการเป็นโรคความดันโลหิตสูงในระเบียนนั้นตรงกันกับ ค่าทำนาย (predicted)

เช่น ระเบียนที่ 69 ค่าจริง (actual)= 1: Negative ค่าทำนาย (predicted) = 1: Negative

จะได้ $y_i = y_{69} = 1$

ระเบียนที่ 81 ค่าจริง (actual)= 1: Positive ค่าทำนาย (predicted)= 1: Positive

จะได้ $y_i = y_{81} = 1$

กำหนดให้ค่า $y_i = 0$ ได้จากกรณีที่ ค่าจริง (actual) ใน class attribute ของข้อมูลการเป็นโรคความดันโลหิตสูงในระเบียนนั้นตรงกันกับ ค่าทำนาย (predicted)

เช่น ระเบียนที่ 68 ค่าจริง (actual)= 2:Positive ค่าทำนาย (predicted)= 1:Negative

จะได้ $y_i = y_{68} = 0$

ระเบียนที่ 87 ค่าจริง (actual)= 1:Negative ค่าทำนาย (predicted)= 2:Positive

จะได้ $y_i = y_{87} = 0$

กำหนดให้ค่า $\hat{y}_i$ ได้จาก ค่าการทำนาย (predicted) ซึ่งอยู่ที่คอลัมภ์ขวาสุดของระเบียนนั้นในช่อง Classifier output

ลำดับที่	y_i	$\hat{y}_i$	e_i^2
1	0	0.999	0.998
2	1	0.999	0.000
3	1	0.999	0.000
4	0	0.999	0.998
5	0	0.999	0.998
6	1	0.999	0.000
7	1	0.999	0.000
8	1	0.999	0.000

ลำดับที่	y_i	$\hat{y}_i$	e_i^2
9	1	0.999	0.000
10	1	0.999	0.000
11	0	0.999	0.998
12	1	0.999	0.000
13	1	0.999	0.000
14	1	0.999	0.000
15	0	0.999	0.998
16	1	0.999	0.000
17	1	0.999	0.000
18	1	0.999	0.000
19	1	0.999	0.000
20	1	0.999	0.000
21	1	0.999	0.000
22	1	0.999	0.000
23	1	0.999	0.000
24	1	0.999	0.000
25	0	0.999	0.998
26	1	0.999	0.000
27	0	0.999	0.998
28	1	0.999	0.000
29	0	0.999	0.998
30	1	0.999	0.000
31	1	0.999	0.000
32	0	0.999	0.998
33	1	0.999	0.000
34	0	0.999	0.998

ลำดับที่	y_i	$\hat{y}_i$	e_i^2
35	1	0.999	0.000
36	1	0.999	0.000
37	1	0.999	0.000
38	1	0.999	0.000
39	0	0.999	0.998
40	1	0.999	0.000
41	1	0.999	0.000
42	1	0.999	0.000
43	1	0.999	0.000
44	1	0.999	0.000
45	1	0.999	0.000
46	0	0.999	0.998
47	1	0.999	0.000
48	1	0.999	0.000
49	1	0.999	0.000
50	1	0.999	0.000
51	1	0.999	0.000
52	0	0.999	0.998
53	0	0.999	0.998
54	1	0.999	0.000
55	1	0.999	0.000
56	1	0.999	0.000
57	1	0.999	0.000
58	1	0.999	0.000
59	1	0.999	0.000
60	1	0.999	0.000

ลำดับที่	y_i	$\hat{y}_i$	e_i^2
61	1	0.999	0.000
62	0	0.999	0.998
63	1	0.999	0.000
64	1	0.999	0.000
65	1	0.999	0.000
66	1	0.999	0.000
67	0	0.999	0.998
68	0	0.999	0.998
69	1	0.999	0.000
70	1	0.999	0.000
71	0	0.999	0.998
72	1	0.999	0.000
73	1	0.999	0.000
74	1	0.999	0.000
75	0	0.999	0.998
76	1	0.999	0.000
77	1	0.999	0.000
78	1	0.999	0.000
79	1	0.999	0.000
80	1	0.999	0.000
81	1	0.999	0.000
82	1	0.999	0.000
83	1	0.999	0.000
84	1	0.999	0.000
85	0	0.999	0.998
86	1	0.999	0.000

ลำดับที่	y_i	$\hat{y}_i$	e_i^2
87	0	0.999	0.998
88	1	0.999	0.000
89	1	0.999	0.000
90	1	0.999	0.000
91	1	0.999	0.000
92	1	0.999	0.000
93	1	0.999	0.000
94	1	0.999	0.000
95	1	0.999	0.000
96	1	0.999	0.000
97	1	0.999	0.000
98	1	0.999	0.000
99	1	0.999	0.000
100	1	0.999	0.000
รวม			20.958

ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) $= \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} = \frac{\sum_{i=1}^{n} e_i^2}{n}$

$$= \frac{20.958}{100} = 0.20958$$

รากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE) $= \sqrt{MSE}$

$$= \sqrt{0.20958} = 0.4578$$

ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยที่หามาได้ เท่ากับ 0.4578 ซึ่งมีค่าใกล้เคียง 0.4576 จาก out put ของโปรแกรม weka