รายงานการวิจัย
การเปรียบเทียบประสิทธิภาพในการทำนายผลการเป็นโรคเบาหวาน
An Efficiency Comparison in Prediction of Diabetes

รศ.สายชล สินสมบูรณ์ทอง

ได้รับทุนสนับสนุนงานวิจัยจากเงินรายได้ ประจำปีงบประมาณ 2559
คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

รายงานการวิจัย
การเปรียบเทียบประสิทธิภาพในการทำนายผลการเป็นโรคเบาหวาน
An Efficiency Comparison in Prediction of Diabetes

รศ.สายชล สินสมบูรณ์ทอง

ได้รับทุนสนับสนุนงานวิจัยจากเงินรายได้ ประจำปีงบประมาณ 2559
คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

หัวข้อ การเปรียบเทียบประสิทธิภาพในการทำนายผลการเป็นโรคเบาหวาน
ผู้วิจัย รศ.สายชล สินสมบูรณ์ทอง
สาขา สถิติประยุกต์
พ.ศ. 2559

บทคัดย่อ

ในการศึกษานี้เป็นการเปรียบเทียบประสิทธิภาพในการทำนายผลการเป็นโรคเบาหวาน 6 วิธี ของโรงพยาบาลแห่งหนึ่ง วิธีการจำแนกกลุ่มที่นำมาเปรียบเทียบ คือ วิธีความใกล้เคียงกันมากที่สุด โดยใช้อัลกอริทึมชนิด IBk วิธีแผนภาพต้นไม้เพื่อการตัดสินใจโดยใช้อัลกอริทึมชนิด J48 วิธีโครงข่ายประสาทโดยใช้อัลกอริทึมชนิดเพอร์เซปตรอนแบบหลายชั้น วิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO ชนิดโพลิโนเมียลเคอร์เนล วิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม และวิธีนาอีฟ เบย์ ในการเปรียบเทียบประสิทธิภาพในการทำนายผลของวิธีการจำแนกกลุ่มทั้ง 6 วิธี จะใช้ค่าความถูกต้อง ค่าอัตราความถูกต้องเชิงบวก ค่าอัตราความถูกต้องเชิงลบ ค่าความแม่นยำ ค่าความระลึก ค่าความถ่วงดุล ค่าอัตราความผิดพลาดเชิงบวก ค่าอัตราความผิดพลาดเชิงลบ ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) และค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE)

ผลการศึกษาพบว่าวิธีโครงข่ายประสาทมีค่าความถูกต้อง ค่าอัตราความถูกต้องเชิงลบ ความแม่นยำ ค่าอัตราความผิดพลาดเชิงบวก ค่าความถ่วงดุล ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย และค่าความคลาดเคลื่อนกำลังสองเฉลี่ยดีที่สุดคือ 95.94%, 0.625, 0.974, 0.375, 0.978, 0.0491 และ 0.0396 ตามลำดับ ส่วนวิธีแผนภาพต้นไม้เพื่อการตัดสินใจและวิธีซัพพอร์ตเวกเตอร์แมชชีนมีค่าอัตราความถูกต้องเชิงบวก ค่าอัตราความผิดพลาดเชิงลบ และค่าความระลึกดีที่สุดคือ 1, 0 และ 1 ตามลำดับ และวิธีนาอีฟ เบย์มีค่าความถูกต้องเชิงลบและความผิดพลาดเชิงบวกดีที่สุดคือ 0.625 และ 0.375 ตามลำดับ เนื่องจากวิธีโครงข่ายประสาทมีประสิทธิภาพในการทำนายผลดีที่สุด 7 ใน 10 ค่า ดังนั้นวิธีโครงข่ายประสาทเป็นวิธีที่ดีที่สุด

คำสำคัญ : วิธีการจำแนกกลุ่ม วิธีความใกล้เคียงกันมากที่สุด วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ วิธีโครงข่ายประสาท วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม และวิธีนาอีฟ เบย์

Thesis Title	An Efficiency Comparison in Prediction of Diabetes
Researcher	Assoc.Prof. Saichon Sinsomboonthong
Programme	Applied Statistics
Year	2016

Abstract

In this study, an efficiency comparison of diabetes prediction were determined and tested against real data obtained from a hospital, and their prediction efficiencies were compared. The tested classification methods were the following: 1) k-nearest neighbor method using IBk algorithm; 2) decision tree method using J48 algorithm; 3) neural network method using multilayer perceptron algorithm; 4) support vector machine method using polynomial kernel; 5) binary logistic regression method; and 6) naïve Bayes method. The following performance values were employed: accuracy, true positive rate, true negative rate, precision, recall, F-Measure, false positive rate, false negative rate, mean absolute error (MAE), and mean square error (MSE).

The important results are as follows. The neural network method showed the best accuracy, true negative rate, precision, false positive rate, F-Measure, MAE, and MSE at 95.94%, 0.625, 0.974, 0.375, 0.978, 0.0491 and 0.0396 respectively. The decision tree method and support vector machine method exhibited the best true positive rate, false negaitive rate and recall at 1, 0 and 1 respectively, while the best true negative and false positive rate at 0.625 and 0.375 were attained by the naïve Bayes method. Since the neural network method offered the best efficiencies for 7 out of the 10 values, it was considered the best prediction method for diabetes.

Keywords: classification method, k-nearest neighbor, decision tree, neural network, support vector machine, binary logistic regression and naïve-Bayes.

กิตติกรรมประกาศ

งานวิจัยเล่มนี้สำเร็จลุล่วงไปได้ด้วยดี เนื่องจากผู้จัดทำได้รับความช่วยเหลือจากบุคคลผู้มีพระคุณหลายท่าน ดังนี้

ขอขอบพระคุณโครงการวิจัยที่เอื้อเฟื้อทุนสนับสนุนในการวิจัยครั้งนี้ โดยใช้เงินรายได้คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ขอขอบคุณนางสาวสุธีรา กิขุนทด และนางสาวขนิษฐา ศรีอุไร นักศึกษาปริญญาตรี ชั้นปีที่ 4 สาขาวิชาสถิติประยุกต์ ภาควิชาสถิติ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ที่คอยให้ความช่วยเหลือด้านการเก็บรวบรวมข้อมูลและวิเคราะห์ข้อมูลบางส่วน

ขอขอบคุณทุกท่านที่มิได้เอ่ยนามในที่นี้ที่ให้คำแนะนำที่เป็นประโยชน์ต่าง ๆ และคอยเป็นกำลังใจให้งานวิจัยเล่มนี้สำเร็จลุล่วงไปได้ด้วยดี

.....

รศ.สายชล สินสมบูรณ์ทอง

(หัวหน้าโครงการวิจัย)

สารบัญ

	หน้า
บทคัดย่อภาษาไทย	I
บทคัดย่อภาษาอังกฤษ	II
กิตติกรรมประกาศ	III
สารบัญ	IV
สารบัญตาราง	VII
สารบัญรูป	XII
บทที่ 1 บทนำ	**1**
1.1 ความสำคัญและที่มาของปัญหา	1
1.2 วัตถุประสงค์ของงานวิจัย	2
1.3 ขอบเขตของงานวิจัย	2
1.4 ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัย	3
1.5 นิยามคำศัพท์	3
บทที่ 2 ทฤษฎีและรายงานวิจัยที่เกี่ยวข้อง	**5**
2.1 ทฤษฎีที่เกี่ยวข้อง	5
2.1.1 วิธีความใกล้เคียงกันมากที่สุด	5
2.1.2 วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ	6
2.1.3 วิธีโครงข่ายประสาท	8
2.1.4 วิธีซัพพอร์ตเวกเตอร์แมชชีน	12
2.1.5 วิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม	16
2.1.6 วิธีนาอีฟ เบย์	16
2.2 รายงานวิจัยที่เกี่ยวข้อง	16
บทที่ 3 วิธีดำเนินงานวิจัย	**28**
3.1 อุปกรณ์ในการวิจัย	28
3.1.1 อุปกรณ์ที่มีอยู่แล้ว	28
3.1.2 อุปกรณ์ที่ต้องการเพิ่มเติม	28
3.2 ขั้นตอนการดำเนินงานวิจัย	28

สารบัญ (ต่อ)

หน้า

บทที่ 4 ผลการวิจัยและอภิปรายผล 32

4.1 การวิเคราะห์ข้อมูล 32

4.1.1 วิธีความใกล้เคียงกันมากที่สุด 32

4.1.1.1 การสร้างตัวแบบการเป็นโรคเบาหวาน 32

4.1.1.2 การทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน 34

4.1.1.3 การทำนายการเป็นโรคเบาหวาน 36

4.1.2 วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ 40

4.1.2.1 การสร้างตัวแบบการเป็นโรคเบาหวาน 40

4.1.2.2 การทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน 40

4.1.2.3 การทำนายการเป็นโรคเบาหวาน 44

4.1.3 วิธีโครงข่ายประสาท 48

4.1.3.1 การสร้างตัวแบบการเป็นโรคเบาหวาน 48

4.1.3.2 การทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน 50

4.1.3.3 การทำนายการเป็นโรคเบาหวาน 51

4.1.4 วิธีซัพพอร์ตเวกเตอร์แมชชีน 61

4.1.4.1 การสร้างตัวแบบการเป็นโรคเบาหวาน 61

4.1.4.2 การทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน 63

4.1.4.3 การทำนายการเป็นโรคเบาหวาน 64

4.1.5 วิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม 69

4.1.5.1 การสร้างตัวแบบการเป็นโรคเบาหวาน 69

4.1.5.2 การทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน 71

4.1.5.3 การทำนายการเป็นโรคเบาหวาน 73

4.1.6 วิธีนาอีฟ เบย์ 78

4.1.6.1 การสร้างตัวแบบการเป็นโรคเบาหวาน 78

4.1.6.2 การทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน 80

4.1.6.3 การทำนายการเป็นโรคเบาหวาน 82

4.2 ผลการเปรียบเทียบประสิทธิภาพในการทำนายผลของวิธีการจำแนกกลุ่ม 86

บทที่ 5 สรุปผลการวิจัยและข้อเสนอแนะ 88

5.1 สรุปผลการวิจัย 88

5.2 การอภิปรายผล 88

5.3 ข้อเสนอแนะ 89

สารบัญ (ต่อ)

ภาคผนวก ก	รายละเอียดและตัวอย่างข้อมูลที่นำมาวิเคราะห์	91
ภาคผนวก ข	การวิเคราะห์ข้อมูล	96
ภาคผนวก ค	ตัวอย่างการคำนวณ	109
เอกสารอ้างอิง		118

สารบัญตาราง

หน้า

ตารางที่ 4.1 ผลในส่วนของการสรุปผลจากการสร้างตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีความใกล้เคียงกันมากที่สุด 32
ตารางที่ 4.2 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการสร้างตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีความใกล้เคียงกันมากที่สุด 33
ตารางที่ 4.3 ผลในส่วนของเมตริกซ์ความสับสนจากการสร้างตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีความใกล้เคียงกันมากที่สุด 33
ตารางที่ 4.4 ผลในส่วนของการสรุปผลจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีความใกล้เคียงกันมากที่สุด 34
ตารางที่ 4.5 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีความใกล้เคียงกันมากที่สุด 34
ตารางที่ 4.6 ผลในส่วนของเมตริกซ์ความสับสนจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีความใกล้เคียงกันมากที่สุด 35
ตารางที่ 4.7 ผลในส่วนของการทำนายชุดข้อมูลทดสอบ ด้วยวิธีความใกล้เคียงกันมากที่สุด 36
ตารางที่ 4.8 ผลในส่วนของการสรุปผลจากการทำนายการเป็นโรคเบาหวาน ด้วยวิธีความใกล้เคียงกันมากที่สุด 38
ตารางที่ 4.9 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการทำนายการเป็นโรคเบาหวาน ด้วยวิธีความใกล้เคียงกันมากที่สุด 39
ตารางที่ 4.10 ผลในส่วนของเมตริกซ์ความสับสนจากการทำนายการเป็นโรคเบาหวาน ด้วยวิธีความใกล้เคียงกันมากที่สุด 39
ตารางที่ 4.11 ผลในส่วนของการสรุปผลจากการสร้างตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ 40
ตารางที่ 4.12 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการสร้างตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ 41
ตารางที่ 4.13 ผลในส่วนของเมตริกซ์ความสับสนจากการสร้างตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ 41
ตารางที่ 4.14 ผลในส่วนของการสรุปผลจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ 42
ตารางที่ 4.15 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ 42

สารบัญตาราง (ต่อ)

หน้า

ตารางที่ 4.16 ผลในส่วนของเมตริกซ์ความสับสนจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ 43
ตารางที่ 4.17 ผลในส่วนของการทำนายชุดข้อมูลทดสอบ ด้วยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ 44
ตารางที่ 4.18 ผลในส่วนของการสรุปผลจากการทำนายการเป็นโรคเบาหวาน ด้วยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ 46
ตารางที่ 4.19 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการทำนายการเป็นโรคเบาหวาน ด้วยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ 47
ตารางที่ 4.20 ผลในส่วนของเมตริกซ์ความสับสนจากการทำนายการเป็นโรคเบาหวาน ด้วยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ 47
ตารางที่ 4.21 ผลในส่วนของการสรุปผลจากการสร้างตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีโครงข่ายประสาท 48
ตารางที่ 4.22 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการสร้างตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีโครงข่ายประสาท 49
ตารางที่ 4.23 ผลในส่วนของเมตริกซ์ความสับสนจากการสร้างตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีโครงข่ายประสาท 49
ตารางที่ 4.24 ผลในส่วนของการสรุปผลจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีโครงข่ายประสาท 50
ตารางที่ 4.25 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีโครงข่ายประสาท 50
ตารางที่ 4.26 ผลในส่วนของเมตริกซ์ความสับสนจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีโครงข่ายประสาท 51
ตารางที่ 4.27 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของการทำนายการเป็นโรคเบาหวาน ด้วยวิธีโครงข่ายประสาท สำหรับอัตราการเรียนรู้ $\eta = 0.1$ และโมเมนตัม $\alpha = 0.5, 0.6, 0.7, 0.8$ และ 0.9 52
ตารางที่ 4.28 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของการทำนายการเป็นโรคเบาหวาน ด้วยวิธีโครงข่ายประสาท สำหรับอัตราการเรียนรู้ $\eta = 0.2$ และโมเมนตัม $\alpha = 0.5, 0.6, 0.7, 0.8$ และ 0.9 53
ตารางที่ 4.29 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของการทำนายการเป็นโรคเบาหวาน ด้วยวิธีโครงข่ายประสาท สำหรับอัตราการเรียนรู้ $\eta = 0.3$ และโมเมนตัม $\alpha = 0.5, 0.6, 0.7, 0.8$ และ 0.9 54

สารบัญตาราง (ต่อ)

หน้า

ตารางที่ 4.30 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของการทำนายการเป็นโรคเบาหวาน ด้วยวิธีโครงข่ายประสาท สำหรับอัตราการเรียนรู้ $\eta = 0.4$ และโมเมนตัม $\alpha = 0.5, 0.6, 0.7, 0.8$ และ 0.9 55
ตารางที่ 4.31 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของการทำนายการเป็นโรคเบาหวาน ด้วยวิธีโครงข่ายประสาท สำหรับอัตราการเรียนรู้ $\eta = 0.5$ และโมเมนตัม $\alpha = 0.5, 0.6, 0.7, 0.8$ และ 0.9 56
ตารางที่ 4.32 ผลในส่วนของการทำนายชุดข้อมูลทดสอบ ด้วยวิธีโครงข่ายประสาท 57
ตารางที่ 4.33 ผลในส่วนของการสรุปผลจากการทำนายการเป็นโรคเบาหวาน ด้วยวิธีโครงข่ายประสาท 59
ตารางที่ 4.34 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการทำนายการเป็นโรคเบาหวาน ด้วยวิธีโครงข่ายประสาท 60
ตารางที่ 4.35 ผลในส่วนของเมตริกซ์ความสับสนจากการทำนายการเป็นโรคเบาหวาน ด้วยวิธีโครงข่ายประสาท 60
ตารางที่ 4.36 ผลในส่วนของการสรุปผลจากการสร้างตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน 61
ตารางที่ 4.37 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการสร้างตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน 62
ตารางที่ 4.38 ผลในส่วนของเมตริกซ์ความสับสนจากการสร้างตัวแบบการพิจารณาการเป็นโรคเบาหวาน ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน 62
ตารางที่ 4.39 ผลในส่วนของการสรุปผลจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน 63
ตารางที่ 4.40 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน 63
ตารางที่ 4.41 ผลในส่วนของเมตริกซ์ความสับสนจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน 64
ตารางที่ 4.42 ผลในส่วนของการทำนายชุดข้อมูลทดสอบ ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน 64
ตารางที่ 4.43 ผลในส่วนของการสรุปผลจากการทำนายการพิจารณาการเป็นโรคเบาหวาน ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน 67
ตารางที่ 4.44 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการทำนายการเป็นโรคเบาหวาน ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน 67

สารบัญตาราง (ต่อ)

หน้า

ตารางที่ 4.45 ผลในส่วนของเมตริกซ์ความสับสนจากการทำนายการเป็นโรคเบาหวาน ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน 68
ตารางที่ 4.46 ผลในส่วนของการสรุปผลจากการสร้างตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม 69
ตารางที่ 4.47 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการสร้างตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม 70
ตารางที่ 4.48 ผลในส่วนของเมตริกซ์ความสับสนจากการสร้างตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม 70
ตารางที่ 4.49 ผลในส่วนของการสรุปผลจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม 71
ตารางที่ 4.50 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม 71
ตารางที่ 4.51 ผลในส่วนของเมตริกซ์ความสับสนจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม 72
ตารางที่ 4.52 ผลในส่วนของการทำนายชุดข้อมูลทดสอบ ด้วยวิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม 73
ตารางที่ 4.53 ผลในส่วนของการสรุปผลจากการทำนายการเป็นโรคเบาหวาน ด้วยวิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม 76
ตารางที่ 4.54 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการทำนายการเป็นโรคเบาหวาน ด้วยวิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม 76
ตารางที่ 4.55 ผลในส่วนของเมตริกซ์ความสับสนจากการทำนายการเป็นโรคเบาหวาน ด้วยวิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม 77
ตารางที่ 4.56 ผลในส่วนของการสรุปผลจากการสร้างตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีนาอีฟ เบย์ 78
ตารางที่ 4.57 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการสร้างตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีนาอีฟ เบย์ 79
ตารางที่ 4.58 ผลในส่วนของเมตริกซ์ความสับสนจากการสร้างตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีนาอีฟ เบย์ 79
ตารางที่ 4.59 ผลในส่วนของการสรุปผลจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีนาอีฟ เบย์ 80
ตารางที่ 4.60 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีนาอีฟ เบย์ 80

สารบัญตาราง (ต่อ)

หน้า

ตารางที่ 4.61 ผลในส่วนของเมตริกซ์ความสับสนจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีนาอีฟ เบย์ 81
ตารางที่ 4.62 ผลในส่วนของการทำนายชุดข้อมูลทดสอบ ด้วยวิธีนาอีฟ เบย์ 82
ตารางที่ 4.63 ผลในส่วนของการสรุปผลจากการทำนายการเป็นโรคเบาหวาน ด้วยวิธีนาอีฟ เบย์ 84
ตารางที่ 4.64 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการทำนายการเป็นโรคเบาหวาน ด้วยวิธีนาอีฟ เบย์ 85
ตารางที่ 4.65 ผลในส่วนของเมตริกซ์ความสับสนจากการทำนายการเป็นโรคเบาหวาน ด้วยวิธีนาอีฟ เบย์ 85
ตารางที่ 4.66 ผลการเปรียบเทียบประสิทธิภาพของวิธีการจำแนกกลุ่มในการเป็นโรคเบาหวานทั้ง 6 วิธี 86
ตารางที่ 4.67 ผลการเปรียบเทียบการทำนายผลของวิธีการจำแนกกลุ่มในการเป็นโรคเบาหวานทั้ง 6 วิธี 87

สารบัญรูป

หน้า

รูปที่ 2.1 ตัวอย่างของวิธีความใกล้เคียงกันมากที่สุด 5
รูปที่ 2.2 ส่วนประกอบของแผนภาพต้นไม้เพื่อการตัดสินใจ 6
รูปที่ 2.3 โครงข่ายของเซลล์ประสาท 9
รูปที่ 2.4 ลักษณะโครงข่ายประสาทเทียมแบบส่งสัญญาณไปข้างหน้า 10
รูปที่ 2.5 ลักษณะโครงข่ายประสาทเทียมแบบมีการย้อนกลับ 11
รูปที่ 2.6 ลักษณะโครงข่ายประสาทเทียมแบบเพอร์เซปตรอนหลายชั้น 11
รูปที่ 2.7 การขยายตัวของเส้นขอบ 13
รูปที่ 2.8 เส้นขอบและเส้นแบ่งเมื่อแทนด้วยสมการเส้นตรง 14
รูปที่ 2.9 รูปแบบการวางตัวที่ไม่สามารถแบ่งด้วยเส้นตรงได้ 15

บทที่ 1
บทนำ

1.1 ความสำคัญและที่มาของปัญหา

เบาหวานเป็นโรคเรื้อรังที่เป็นปัญหาสำคัญทางด้านสาธารณสุขของโลก เป็นภัยคุกคามที่ลุกลามอย่างรวดเร็วไปทั่วโลก ส่งผลกระทบต่อการพัฒนาทางเศรษฐกิจอย่างมาก จากข้อมูลสมาพันธ์เบาหวานนานาชาติ (International Diabetes Federation : IDF) ได้รายงานว่าในปัจจุบันทั่วโลกมีผู้เสียชีวิตด้วยโรคเบาหวาน 4 ล้านคนต่อปี เฉลี่ย 8 วินาทีต่อ 1 คน สำหรับผู้เป็นเบาหวานพบมากกว่า 300 ล้านคนและพบว่าคนที่อยู่ในประเทศที่มีรายได้ต่ำและปานกลางมีโอกาสเป็นเบาหวานเร็วกว่าคนที่อยู่ในประเทศที่มีรายได้สูง 10-20 ปี โดยพบมากขึ้นในวัยทำงาน (นุชรี อาบสุวรรณ และนิตยา พันธุเวทย์, 2554)

ปัจจัยเสี่ยงของการเกิดโรคเบาหวานคือ น้ำหนักมากเกินไปและอ้วน ขาดการออกกำลังกาย เคยตรวจพบน้ำตาล ในเลือดสูงกว่าปกติ การรับประทานอาหารไม่ถูกสัดส่วน หวาน มัน เค็มมากเกินไป อายุที่เพิ่มขึ้น ความดันโลหิตสูง และไขมันในเลือดสูง มีประวัติญาติสายตรงเป็นโรคเบาหวาน มีประวัติเบาหวานขณะตั้งครรภ์ และพันธุกรรม (http:/www.ifd.org/worlddiabetesday/act-on-diabetes-now)

ปีพ.ศ.2555 สำนักระบาดวิทยาได้รับข้อมูลแฟ้มมาตรฐานจากสำนักนโยบายและยุทธศาสตร์ สำนักงานปลัดกระทรวง กระทรวงสาธารณสุข ซึ่งรับข้อมูลจากสำนักงานสาธารณสุขจังหวัดทั่วประเทศ จำนวน 76 จังหวัด ยกเว้นกรุงเทพมหานคร พบว่าผู้ป่วยสะสม ณ วันที่ 31 ธันวาคม พ.ศ. 2555 จำนวน 1,799,977 ราย ความชุก 2,800.80 ต่อประชากรแสนคน และระหว่างวันที่ 1 มกราคม พ.ศ. 2555 – 31 ธันวาคม พ.ศ. 2555 มีผู้ป่วยเบาหวานที่ได้รับการขึ้นทะเบียนเป็นผู้ป่วยรายใหม่จำนวน 336,265 ราย อุบัติการณ์ 523.24 ต่อประชากรแสนคน โดยมีแนวโน้มเพิ่มขึ้นอย่างต่อเนื่องในช่วงปี พ.ศ. 2551-2555

ผู้ป่วยเบาหวานที่ได้รับการขึ้นทะเบียนเป็นผู้ป่วยรายใหม่เพศชาย 116,715 ราย เพศหญิง 219,550 ราย สัดส่วนเพศชายต่อเพศหญิงเท่ากับ 1:1.9 กลุ่มอายุที่พบสูงสุดคือกลุ่มอายุมากกว่า 60 ปี รองลงมาคือกลุ่มอายุ 50-59 ปี 40-49 ปี 15-39 ปี และน้อยกว่า 15 ปี โดยมีอัตราป่วยเท่ากับ 2,128.04, 1,207.35, 463.44, 61.61 และ 8.29 ต่อประชากรแสนคน ตามลำดับ

ภาคที่ได้รับรายงานมากที่สุดคือภาคกลาง รองลงมาคือภาคตะวันออกเฉียงเหนือ ภาคเหนือ และภาคใต้ โดยมีอัตราป่วยเท่ากับ 643.16, 603.46, 521.66, 447.33 ต่อประชากรแสนคน ตามลำดับ จังหวัดที่มีอุบัติการณ์ต่อประชากรแสนคนสูงสุด 10 จังหวัด คือ จังหวัดพิจิตร รองลงมา จังหวัดปราจีนบุรี มหาสารคาม สุราษฎร์ธานี เลย จันทบุรี น่าน บึงกาฬ นครปฐม และอ่างทอง ผู้ป่วยส่วนใหญ่เข้ารับการรักษาที่โรงพยาบาลส่งเสริมสุขภาพตำบล ร้อยละ 59.38 รองลงมาโรงพยาบาลชุมชน ร้อยละ 24.94 โรงพยาบาลศูนย์/โรงพยาบาลทั่วไป ร้อยละ 9.94 และอื่น ๆ ร้อยละ 5.74 (อมรา ทองหงษ์, 2555)

ด้วยสาเหตุดังกล่าวข้างต้น ผู้วิจัยจึงให้ความสนใจในการเปรียบเทียบประสิทธิภาพในการทำนายผลการเป็นโรคเบาหวานของผู้มาตรวจร่างกายในโรงพยาบาลซึ่งการเปรียบเทียบประสิทธิภาพในการทำนายผลการเป็นโรคเบาหวานขึ้นอยู่กับปัจจัยหลายอย่าง เช่น เพศ อายุ น้ำหนัก ส่วนสูง

ค่าดัชนีมวลกาย การออกกำลังกาย ความดันโลหิตสูง ความเครียด การดื่มเครื่องดื่มที่มีแอลกอฮอล์ กรรมพันธุ์ การรับประทานอาหาร เป็นต้น โดยผู้วิจัยสนใจศึกษาการเปรียบเทียบประสิทธิภาพในการทำนายผลการเป็นโรคเบาหวานของผู้มาตรวจร่างกายที่โรงพยาบาล 6 วิธี คือ วิธีความใกล้เคียงกันมากที่สุด วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ วิธีโครงข่ายประสาท วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีการถดถอยโลจิสติก และวิธีนาอีฟ เบย์ เพื่อหาวิธีที่มีประสิทธิภาพและทำนายผลการเป็นโรคเบาหวานที่เหมาะสมต่อไป

1.2 วัตถุประสงค์ของงานวิจัย

เพื่อเปรียบเทียบประสิทธิภาพในการทำนายผลการเป็นโรคเบาหวาน ระหว่างวิธีความใกล้เคียงกันมากที่สุด วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ วิธีโครงข่ายประสาท วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม และวิธีนาอีฟ เบย์

1.3 ขอบเขตของงานวิจัย

โครงการวิจัยนี้ได้ทำการศึกษาเรื่องการเปรียบเทียบประสิทธิภาพในการทำนายผลการเป็นโรคเบาหวาน 6 วิธี คือ วิธีความใกล้เคียงกันมากที่สุด โดยใช้อัลกอริทึมชนิด IBk วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ โดยใช้อัลกอริทึมชนิด J48 วิธีโครงข่ายประสาท โดยใช้อัลกอริทึมชนิดเพอร์เซปตรอนแบบหลายชั้น โดยกำหนดค่าอัตราการเรียนรู้เป็น 0.1, 0.2, 0.3, 0.4 และ 0.5 ค่าโมเมนตัมเป็น 0.5, 0.6, 0.7, 0.8 และ 0.9 จำนวนรอบการสอน 20,000 รอบ และชั้นซ่อน 1 ชั้น วิธีซัพพอร์ตเวกเตอร์แมชชีน ใช้อัลกอริทึม SMO ชนิดโพลิโนเมียลเคอร์เนล วิธีการถดถอยโลจิสติก และวิธีนาอีฟเบย์ ในการเปรียบเทียบประสิทธิภาพของวิธีการจำแนกกลุ่มทั้ง 6 วิธี ใช้ค่าความถูกต้อง (Accuracy) ค่าอัตราความถูกต้องเชิงบวก (TP Rate) ค่าอัตราความถูกต้องเชิงลบ (TN Rate) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าความถ่วงดุล (F-Measure) โดยพิจารณาจากค่าเหล่านี้ที่มีค่ามากที่สุด ส่วนค่าอัตราความผิดพลาดเชิงบวก (FP Rate) ค่าอัตราความผิดพลาดเชิงลบ (FN Rate) ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Error) และค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error : MSE) โดยพิจารณาจากค่าเหล่านี้ที่มีค่าน้อยที่สุดจะทำให้มีประสิทธิภาพในการทำนายผลดีที่สุด ทำการเก็บรวบรวมข้อมูลจากผู้มาตรวจร่างกายการเป็นโรคเบาหวานในโรงพยาบาลตั้งแต่เดือนมกราคมถึงเดือนเมษายน พ.ศ. 2559 จำนวนประมาณ 2000 คน ทำการเก็บข้อมูลเกี่ยวกับตัวแปรอิสระซึ่งประกอบด้วยเพศ อายุ น้ำหนัก ส่วนสูง ค่าดัชนีมวลกาย การออกกำลังกาย ความดันโลหิตสูง ความเครียด การดื่มเครื่องดื่มที่มีแอลกอฮอล์ กรรมพันธุ์ การรับประทานอาหาร และตัวแปรตามประกอบด้วย ผลการตรวจวินิจฉัยการเป็นโรคเบาหวานของผู้มาตรวจร่างกาย โดยแบ่งเป็น 2 กลุ่ม คือ 1) กลุ่มที่ไม่เป็นโรคเบาหวาน และ 2) กลุ่มที่เป็นโรคเบาหวาน รวมระยะเวลาดำเนินโครงการ 1 ปี

1.4 ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัย

1.4.1 สามารถนำผลการเปรียบเทียบประสิทธิภาพในการทำนายผลการเป็นโรคเบาหวานไปใช้เป็นแนวทางในการเลือกวิธีจำแนกกลุ่มที่เหมาะสม

1.4.2 สามารถนำผลการเปรียบเทียบประสิทธิภาพในการทำนายผลการเป็นโรคเบาหวานไปใช้เป็นแนวทางในการทำนายผลการเป็นโรคเบาหวานของผู้มาตรวจร่ายกายในโรงพยาบาล

1.5 นิยามคำศัพท์

โรคเบาหวาน (Diabetes) เป็นโรคที่เกิดขึ้นเมื่อร่างกายผลิตอินซูลินไม่เพียงพอหรือร่างกายไม่สามารถใช้อินซูลินได้อย่างเหมาะสม ทำให้ปริมาณกลูโคสในเลือดสูง เซลล์ในร่างกายย่อยอาหารได้เป็นกลูโคสและกลูโคสเหล่านี้ถูกขนส่งไปยังเซลล์ต่าง ๆ ในร่างกาย อินซูลินเป็นฮอร์โมนซึ่งผลิตกลูโคสโดยตรงโดยทำการย่อยอาหารในเซลล์ร่างกาย การเปลี่ยนแปลงในการผลิตอินซูลินบางอย่างจะทำให้ระดับน้ำตาลในเลือดเพิ่มขึ้น ส่งผลเสียต่อเนื้อเยื่อและความล้มเหลวของอวัยวะ คนโดยทั่วไปได้รับความทุกข์ทรมานจากโรคเบาหวานเมื่อระดับน้ำตาลในเลือดสูงกว่าปกติ (4.4 ถึง 6.1 mmol/L) (Iyer, A. et al. : 2015)

วิธีความใกล้เคียงกันมากที่สุด (K-Nearest Neighbor (KNN) Method) เป็นวิธีที่ไม่มีการสร้างตัวแบบจากข้อมูลฝึกหัดเก็บไว้ ทำนายข้อมูลใหม่โดยอาศัยการเปรียบเทียบกับข้อมูลฝึกหัดจำนวน k ตัว ที่อยู่ใกล้เคียงกันมากที่สุด ใช้คำตอบของข้อมูลฝึกหัดที่อยู่ใกล้เคียงกันมากที่สุด k ตัว ที่พบมากที่สุดเป็นคำตอบ วิธีนี้ทำนายได้เฉพาะข้อมูลเชิงกลุ่ม (nominal data) เท่านั้น

วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ (Decision Tree Method) ป็นตัวแบบทางคณิตศาสตร์เพื่อหาทางเลือกที่ดีที่สุดโดยการนำข้อมูลมาสร้างตัวแบบการพยากรณ์ในรูปแบบของโครงสร้างต้นไม้ซึ่งมีการเรียนรู้ข้อมูลแบบมีผู้สอน (supervised learning) สามารถสร้างตัวแบบการจำแนกกลุ่มได้จากกลุ่มตัวอย่างของข้อมูลฝึกหัดโดยอัตโนมัติและสามารถพยากรณ์กลุ่มของรายการที่ยังไม่เคยนำมาจำแนกกลุ่มได้อีกด้วย

วิธีโครงข่ายประสาท (Neural Network Method) เป็นวิธีที่ใช้หลักการเลียนแบบการทำงานของสมองมนุษย์ เส้นเชื่อมแต่ละเส้นจะมีน้ำหนักถ่วง (weight) เพื่อใช้กำหนดน้ำหนักถ่วงหรือความสำคัญของข้อมูลเข้า (input data) กำหนดค่าเริ่มต้นโดยการสุ่ม ในแต่ละโหนดทำการคำนวณค่าผลรวมเชิงเส้นแบบถ่วงน้ำหนักและผ่านฟังก์ชันกระตุ้น (activation function) คำนวณค่าความคลาดเคลื่อน (error) ระหว่างคำตอบที่ทำนายได้กับเฉลย ถ้ามีความคลาดเคลื่อนเกิดขึ้น ระบบจะทำการปรับปรุงค่าน้ำหนักถ่วงของแต่ละการเชื่อมต่อ (connection) ทำนายข้อมูลได้ทั้งข้อมูลเชิงกลุ่ม (nominal data) และข้อมูลเชิงตัวเลข (numeric data) วิธีโครงข่ายประสาทอยู่ในหมวดวิธีที่เป็น Functions ใช้สมการทางคณิตศาสตร์ในการสร้างตัวแบบ

วิธีซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine Method) เป็นวิธีที่ใช้กระบวนการสอนเครื่องแบบมีผู้สอน (supervised learning) เพื่อให้สามารถสร้างตัวจำแนกข้อมูล (classifier) ที่มีความทั่วไป (generalize) สูง นั่นคือสามารถทำงานได้ดีกับตัวอย่างที่ไม่รู้จัก (unknown database) ด้วยกระบวนการปรับรูปแบบข้อมูลจากข้อมูลที่มีมิติต่ำ (low dimension dataset) บนพื้นที่ข้อมูลนำเข้า (input space) ให้อยู่ในรูปแบบของข้อมูลที่มีมิติสูง (high dimension dataset) บนพื้นที่ข้อมูลคุณลักษณะ (feature space) โดยใช้ฟังก์ชันในการ

ปรับรูปแบบข้อมูลเรียกว่า**ฟังก์ชันเคอร์เนล (kernel function)** ซึ่งความสามารถดังกล่าวช่วยให้การสร้างตัวจำแนกข้อมูลด้วยสมการกำลังสอง (quadratic equation) บนพื้นที่ข้อมูลคุณลักษณะเป็นไปได้ง่ายขึ้นและมีความชัดเจนในการจำแนกกลุ่มมากยิ่งขึ้นด้วย

วิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม (Binary Logistic Regression Method) เป็นวิธีการวิเคราะห์การถดถอยแบบหนึ่งโดยที่ตัวแปรตามเป็นตัวแปรเชิงคุณภาพที่มีค่าได้เพียง 2 ค่า (dichotomous or binary variable) ส่วนตัวแปรอิสระอาจจะเป็นตัวแปรเชิงปริมาณหรือเชิงคุณภาพ หรืออาจจะมีทั้งตัวแปรเชิงปริมาณและตัวแปรเชิงคุณภาพก็ได้

วิธีนาอีฟ เบย์ (Naïve Bayes Method) คืออัลกอริทึมที่ใช้หลักการของความน่าจะเป็นในการคัดกรองแต่ละคำตอบ (Class) โดยมีคำตอบ 2 คำตอบ

บทที่ 2
ทฤษฎีและรายงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 วิธีความใกล้เคียงกันมากที่สุด (K-Nearest Neighbor (KNN) Method)

ไม่มีการสร้างตัวแบบจากข้อมูลฝึกหัดเก็บไว้ ทำนายข้อมูลใหม่โดยอาศัยการเปรียบเทียบกับข้อมูลฝึกหัดจำนวน k ตัว ที่อยู่ใกล้เคียงกันมากที่สุด ใช้คำตอบของข้อมูลฝึกหัดที่อยู่ใกล้เคียงกันมากที่สุด k ตัว ที่พบมากที่สุดเป็นคำตอบ วิธีนี้ทำนายได้เฉพาะข้อมูลเชิงกลุ่ม (nominal data) เท่านั้น

วิธีความใกล้เคียงกันมากที่สุดเป็นวิธีการที่ได้รับความนิยมในการใช้งานอย่างมาก เนื่องจากเป็นวิธีการที่ง่ายและมีประสิทธิภาพซึ่งสามารถนำไปประยุกต์ใช้กับงานได้หลายอย่าง เช่น งานทางด้านการจำแนกกลุ่ม รวมถึงงานทางด้านการแทนที่ข้อมูลสูญหาย ซึ่งมีขั้นตอนการดำเนินการดังนี้ (Tan, P-N. and et al. : 2006)

2.1.1.1 กำหนดค่า k เพื่อใช้พิจารณาสมาชิกที่อยู่ใกล้เคียงกันมากที่สุด เช่น k = 3 คือ จะพิจารณาเฉพาะข้อมูล 3 ตัวแรกที่อยู่ใกล้กับจุดที่ต้องการจะทำนาย

2.1.1.2 คำนวณหาระยะห่างระหว่างข้อมูลตัวอย่างที่สนใจกับข้อมูลอื่น ๆ ทุกตัวด้วยวิธีระยะห่างยูคลิเดียน (euclidean distance) จากสมการดังนี้

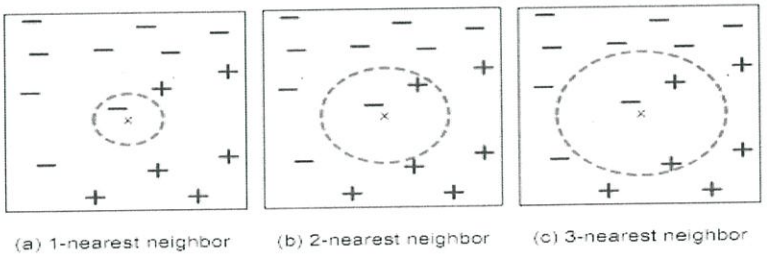
$$\text{dist}(x_i, x_j) = \sqrt{\sum_{k=1}^{n}(x_{i,k} - x_{j,k})^2} \qquad (2.1)$$

โดยที่ $\text{dist}(x_i, x_j)$ คือ ระยะห่างระหว่างตัวอย่าง x_i กับตัวอย่าง x_j

n คือ จำนวนคุณสมบัติทั้งหมดของตัวอย่าง

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

2.1.1.3 เลือกค่าข้อมูลที่มีค่าระยะห่างน้อยที่สุด k ตัว เพื่อนำมาพิจารณาหาคำตอบ ดังรูปที่ 2.1



(a) ความใกล้เคียงกันมากที่สุดโดยพิจารณาจากข้อมูล 1 ตัว

(b) ความใกล้เคียงกันมากที่สุดโดยพิจารณาจากข้อมูล 2 ตัว

(c) ความใกล้เคียงกันมากที่สุดโดยพิจารณาจากข้อมูล 3 ตัว

รูปที่ 2.1 ตัวอย่างของวิธีความใกล้เคียงกันมากที่สุด

2.1.2 วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ (Decision Tree Method)

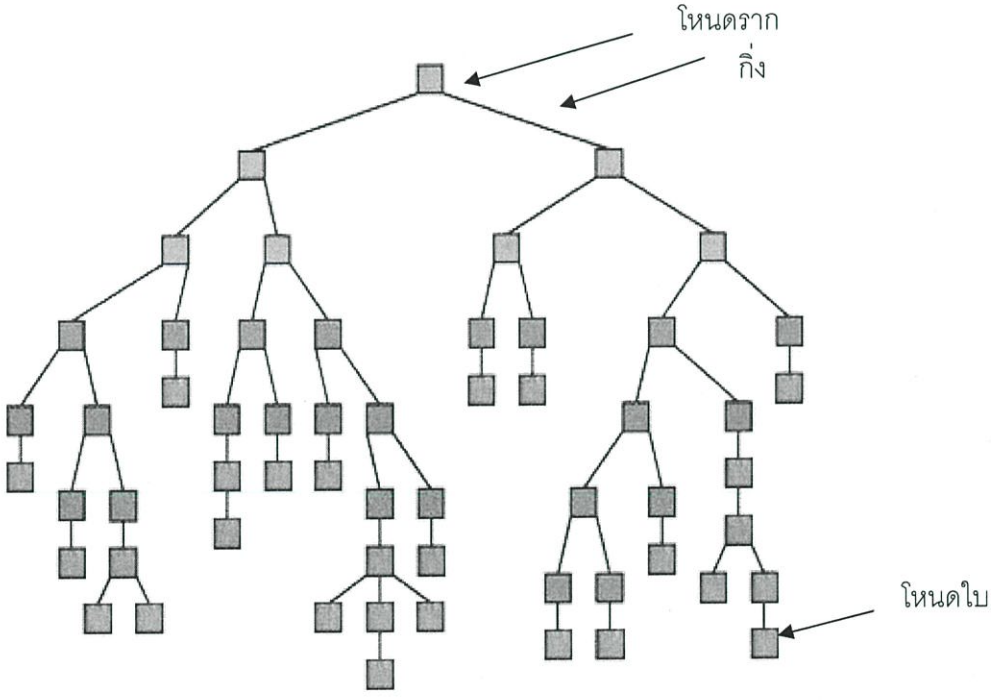
วิธีแผนภาพต้นไม้เพื่อการตัดสินใจเป็นตัวแบบทางคณิตศาสตร์เพื่อหาทางเลือกที่ดีที่สุดโดยการนำข้อมูลมาสร้างตัวแบบการพยากรณ์ในรูปแบบของโครงสร้างต้นไม้ซึ่งมีการเรียนรู้ข้อมูลแบบมีผู้สอน (supervised learning) สามารถสร้างตัวแบบการจำแนกกลุ่มได้จากกลุ่มตัวอย่างของข้อมูลฝึกหัดโดยอัตโนมัติและสามารถพยากรณ์กลุ่มของรายการที่ยังไม่เคยนำมาจำแนกกลุ่มได้อีกด้วย

2.1.2.1 ส่วนประกอบของแผนภาพต้นไม้เพื่อการตัดสินใจ

1) โหนด (Node) คือคุณสมบัติต่าง ๆ เป็นจุดที่แยกข้อมูลว่าจะให้ไปในทิศทางใด ซึ่งโหนดที่อยู่สูงสุดเรียกว่า โหนดราก (root node)

2) กิ่ง (Branch) คือคุณสมบัติของโหนดที่แตกออกมาโดยจำนวนของกิ่งจะเท่ากับคุณสมบัติของโหนด

3) ใบ (Leaf) คือกลุ่มของผลลัพธ์ในการแยกแยะข้อมูล ซึ่งโหนดที่อยู่ล่างสุดเรียกว่า โหนดใบ (leaf node) โดยสามารถแสดงส่วนประกอบของแผนภาพต้นไม้เพื่อการตัดสินใจดังรูปที่ 2.2



รูปที่ 2.2 ส่วนประกอบของแผนภาพต้นไม้เพื่อการตัดสินใจ

2.1.2.2 การสร้างแผนภาพต้นไม้เพื่อการตัดสินใจ

หลักการพื้นฐานของการสร้างแผนภาพต้นไม้เพื่อการตัดสินใจเป็นการสร้างจากบนลงล่าง คือ เริ่มจากการสร้างรากของต้นไม้ก่อน แล้วจึงแตกกิ่งไปจนถึงใบ โดยแสดงขั้นตอนการสร้างแผนภาพต้นไม้เพื่อการตัดสินใจได้ดังนี้

1) ต้นไม้เริ่มต้นโดยมีโหนดเพียงโหนดเดียวแสดงถึงชุดข้อมูลฝึกหัด (training data set)

2) ถ้าข้อมูลทั้งหมดอยู่ในกลุ่มเดียวกันแล้ว ให้โหนดนั้นเป็นใบและตั้งชื่อแยกตามกลุ่มของข้อมูลนั้น

3) ถ้าในโหนดมีข้อมูลหลายกลุ่มปะปนอยู่ จะต้องวัดค่าผลกำไร (gain) ของ แต่ละคุณลักษณะ (attribute) เพื่อที่จะใช้เป็นเกณฑ์ (criterion) ในการคัดเลือกคุณลักษณะที่มีความสามารถในการแบ่งข้อมูลออกเป็นกลุ่มต่าง ๆ ได้ดีที่สุด โดยคุณลักษณะที่มีผลกำไรมากที่สุดจะถูกเลือกให้เป็นตัวทดสอบหรือคุณลักษณะที่ใช้ในการตัดสินใจโดยแสดงในรูปของโหนดบนต้นไม้

4) กิ่งของต้นไม้ถูกสร้างขึ้นจากค่าต่าง ๆ ที่เป็นไปได้ของโหนดทดสอบและข้อมูลจะถูกแบ่งออกตามกิ่งต่าง ๆ ที่สร้างขึ้น

5) ทำการวนซ้ำเพื่อหาคุณลักษณะที่มีผลกำไรมากที่สุด สำหรับข้อมูลที่ถูกแบ่งแยกออกมาในแต่ละกิ่งเพื่อนำคุณลักษณะนี้มาสร้างเป็นโหนดตัดสินใจต่อไป โดยที่คุณลักษณะที่ถูกเลือกมาเป็นโหนดแล้วจะไม่ถูกเลือกมาอีกสำหรับโหนดในระดับต่อ ๆ ไป

6) ทำการวนซ้ำเพื่อแบ่งข้อมูลและแตกกิ่งของต้นไม้ไปเรื่อย ๆ โดยการวนซ้ำจะสิ้นสุดก็ต่อเมื่อเงื่อนไขข้อใดข้อหนึ่งข้างบนนี้เป็นจริง

2.1.2.3 การคำนวณค่าผลกำไรสารสนเทศ (Information Gain)

แผนภาพต้นไม้เพื่อการตัดสินใจเป็นโครงสร้างที่ใช้แสดงกฎที่ได้จากเทคนิคการจำแนกกลุ่มของข้อมูล โดยแผนภาพต้นไม้เพื่อการตัดสินใจจะมีลักษณะคล้ายโครงสร้างต้นไม้ โดยที่แต่ละโหนดแสดงคุณลักษณะ ในการสร้างแผนภาพต้นไม้เพื่อการตัดสินใจ ปัญหาที่สำคัญที่ต้องพิจารณาคือควรจะตัดสินใจเลือกคุณลักษณะใดมาทำหน้าที่เป็นโหนดราก ในแต่ละขั้นตอนของการสร้างต้นไม้และต้นไม้ย่อย (subtree) ของแผนภาพต้นไม้ เพื่อการตัดสินใจ เกณฑ์ที่ใช้ช่วยประกอบการเลือกคุณลักษณะเพื่อการคำนวณเกณฑ์ผลกำไร (gain criterion) ซึ่งเป็นค่าที่บ่งบอกว่าคุณลักษณะที่เป็นไปได้จากชุดข้อมูลมาทำหน้าที่เป็นโหนดราก ถ้าคุณลักษณะใดให้ผลกำไรสูงที่สุดแสดงว่าคุณลักษณะนั้นสามารถจำแนกกลุ่มของข้อมูล ได้ดีที่สุด การใช้ผลกำไรสารสนเทศจะช่วยลดจำนวนครั้งของการทดสอบในการแยกแยะข้อมูล อีกทั้งยังรับประกันว่าแผนภาพต้นไม้เพื่อการตัดสินใจที่ได้จะไม่มีความซับซ้อนมากเกินไป (overfitting) ซึ่งผลกำไรสารสนเทศนั้นสามารถคำนวณได้ดังนี้

$$I(S_1, S_2, \ldots, S_n) = -\sum_{i=1}^{n} \frac{S_i}{S} \log_2 \frac{S_i}{S} \qquad (2.2)$$

เมื่อ S คือ เซตของข้อมูลซึ่งประกอบด้วยข้อมูล S ระเบียน (record)

n คือ จำนวนกลุ่มทั้งหมดที่ต่างกันของข้อมูลชุดนั้น

S_i คือ จำนวนข้อมูลที่เป็นสมาชิกของ S และอยู่ในกลุ่ม C_i

C_i คือ กลุ่มในลำดับที่ i โดยที่ i มีค่าระหว่าง 1 ถึง n

ค่าเอ็นโทรปี (Entropy) ของคุณลักษณะ A ซึ่งมีค่าของคุณลักษณะเป็น $(a_1, a_2, \ldots, a_v)$ หาได้ดังนี้

$$E(A) = \sum_{j=1}^{v} \frac{S_{1j} + \ldots + S_{nj}}{S} I(S_{1j} + \ldots + S_{nj}) \qquad (2.3)$$

S_{ij} คือ จำนวนข้อมูลที่เป็นสมาชิกของ S และอยู่ในกลุ่ม C_i จากการแบ่งข้อมูลด้วยค่าที่เป็นไปได้ของคุณลักษณะ A

ดังนั้นจะสามารถพิจารณาเกณฑ์ผลกำไรได้ดังนี้

$$\text{Gain}(A) = I(S_{1j}, S_{2j}, \ldots, S_{nj}) - E(A) \quad (2.4)$$

ข้อดี

1) เข้าใจได้ง่าย
2) สร้างกฎได้จากต้นไม้
3) เลือกเฉพาะคุณลักษณะ (attribute) ที่สำคัญในการสร้างตัวแบบ

ข้อเสีย

1) ใช้ได้กับคำตอบ (class) ที่เป็นข้อมูลเชิงกลุ่ม (nominal data) เท่านั้น
2) ความถูกต้องในการทำนายไม่สูง

การประยุกต์ใช้แผนภาพต้นไม้เพื่อการตัดสินใจ (Decision tree application)

แผนภาพต้นไม้เพื่อการตัดสินใจใช้ตอบคำถามที่ต้องการจำแนกประเภทข้อมูลที่ต้องการความเข้าใจประกอบ

ใช้ในการพิจารณาให้สินเชื่อแก่บุคลต่าง ๆ

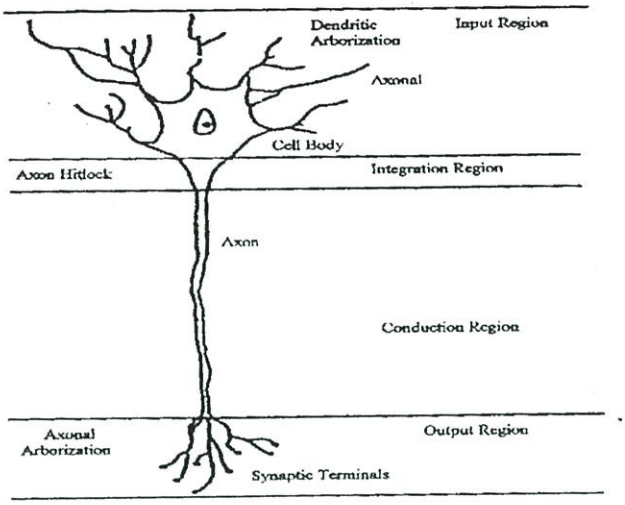
ใช้ในการทำนายว่าลูกค้าคนไหนที่มีโอกาสจะยกเลิกการใช้บริการและเหตุผลเพราะอะไร

2.1.3 วิธีโครงข่ายประสาท (Neural Network Method)

วิธีโครงข่ายประสาทใช้หลักการเลียนแบบการทำงานของสมองมนุษย์ เส้นเชื่อมแต่ละเส้น จะมีน้ำหนักถ่วง (weight) เพื่อใช้กำหนดน้ำหนักถ่วงหรือความสำคัญของข้อมูลเข้า (input data) กำหนดค่าเริ่มต้นโดยการสุ่ม ในแต่ละโหนดทำการคำนวณค่าผลรวมเชิงเส้นแบบถ่วงน้ำหนักและผ่านฟังก์ชันกระตุ้น (activation function) คำนวณค่าความคลาดเคลื่อน (error) ระหว่างคำตอบที่ทำนายได้กับเฉลย ถ้ามีความคลาดเคลื่อนเกิดขึ้น ระบบจะทำการปรับปรุงค่าน้ำหนักถ่วงของแต่ละการเชื่อมต่อ (connection) ทำนายข้อมูลได้ทั้งข้อมูลเชิงกลุ่ม (nominal data) และข้อมูลเชิงตัวเลข (numeric data) วิธีโครงข่ายประสาทอยู่ในหมวดวิธีที่เป็น Functions ใช้สมการทางคณิตศาสตร์ในการสร้างตัวแบบ

โครงข่ายประสาทเป็นศาสตร์ที่จำลองแบบความสามารถของมนุษย์ด้านการเรียนรู้ จดจำและจำแนกสิ่งต่าง ๆ ซึ่งใช้สมองเป็นส่วนสำคัญ ในการประมวลระบบของโครงข่ายประสาทนั้นจะเลียนแบบการทำงานของระบบสมองคือมีการส่งผ่านข้อมูลระหว่างกันโดยมีการเชื่อมต่อของเซลล์ประสาท (neuron) กันเป็นโครงข่ายร่างแหจำนวนมากและมีการประมวลผลในลักษณะขนาน (parallel processing) สาเหตุหลักที่โครงข่ายประสาทเป็นที่นิยมมากขึ้นเนื่องจากมีความยืดหยุ่นในการทำงานสูงและสามารถปรับตัวเองให้ทำงานในสภาพที่เปลี่ยนแปลงได้ดี อีกทั้งยังไม่จำเป็นต้องทราบตัวแบบทางคณิตศาสตร์ที่แน่นอนของกระบวนการ เพียงแต่ใช้ชุดข้อมูลที่ประกอบด้วยข้อมูลเข้า (input data) และข้อมูลเป้าหมาย (target data) ของกระบวนการในจำนวนที่มากพอมาใช้ในการสอนโครงข่ายประสาท

2.1.3.1 ความรู้พื้นฐานของระบบประสาท (Neural system knowledge)



รูปที่ 2.3 โครงข่ายของเซลล์ประสาท

ภายในสมองมนุษย์ประกอบด้วยหน่วยประมวลผลขนาดเล็ก เรียกว่า เซลล์ประสาท (neuron) ซึ่งจะมีประมาณ 10 หน่วย ในเซลล์ประสาทแต่ละหน่วยดังแสดงในรูปที่ 2.3 ประกอบด้วยใยประสาท (dendrites) ตัวเซลล์ (cell body) และเส้นใยประสาท (axon) ซึ่งแบ่งออกเป็น 4 บริเวณคือ

1) บริเวณนำกระแสประสาทเข้า (Input region) เป็นบริเวณที่จะมีการนำกระแสประสาท (nerve impulse) จากเซลล์ประสาทอื่นเข้ามาภายในตัวเซลล์โดยผ่านทางใยประสาทซึ่งมีลักษณะแตกเป็นกิ่งก้านคล้ายต้นไม้และมีจำนวนตั้งแต่ 1 ใยขึ้นไป

2) บริเวณการรวมกระแสประสาทเข้า (Integration region) เป็นบริเวณที่มีการรวมกระแสประสาทก่อนที่จะเข้าสู่บริเวณการนำแสประสาทรวมออกจากเซลล์

3) บริเวณการนำกระแสประสาทรวมออกจากเซลล์ (Conduction region) เป็นบริเวณที่จะนำกระแสประสาทรวมออกจากเซลล์โดยใช้เส้นใยประสาทเป็นทางผ่านซึ่งมีเพียง 1 เส้นใยต่อเซลล์เท่านั้น

4) บริเวณการนำกระแสประสาทรวมออก (Output region) เป็นบริเวณส่วนปลายของเส้นใยประสาทที่มีการแตกแขนงใช้ในการถ่ายทอดกระแสประสาทข้ามเซลล์ไปยังเซลล์ประสาทอื่นโดยผ่านทางใยประสาทของเซลล์ประสาทนั้น

2.1.3.2 การเรียนรู้ของโครงข่ายประสาท (Neural network learning)

การเรียนรู้ของโครงข่ายประสาทจะมีประสิทธิภาพเพียงใดนั้นขึ้นอยู่กับค่าถ่วงน้ำหนัก (weight) ของโครงข่ายที่ทำการออกแบบ ซึ่งการฝึกหัดโครงข่ายประสาทคือการหาค่าถ่วงน้ำหนักที่เหมาะสมให้กับโครงข่ายประสาทนั้น ๆ โดยทั่วไปสามารถจำแนกวิธีการเรียนรู้ของโครงข่ายประสาทได้เป็น 2 ประเภท คือ การเรียนรู้แบบมีผู้สอนและการเรียนรู้แบบไม่มีผู้สอน

1) การเรียนรู้แบบมีผู้สอน (Supervised learning)

การเรียนรู้แบบมีผู้สอนจะกำหนดข้อมูลฝึกหัด (training data set) ให้กับโครงข่ายประสาท ซึ่งกลุ่มนี้ประกอบด้วยข้อมูลเข้า (Input data) และข้อมูลเป้าหมาย (target data) ที่ต้องการ จากนั้นโครงข่ายประสาทจะทำการคำนวณค่าถ่วงน้ำหนักที่เหมาะสมให้กับข้อมูลฝึกหัด โดยคำตอบที่ได้จากโครงข่ายประสาทจะถูกคำนวณค่าความผิดพลาด (error value) ว่ามีความห่างจากคำตอบที่ต้องการของข้อมูลนำเข้าในชุดเดียวกันมากน้อยเพียงใด ถ้ายังมีความผิดพลาดสูงอยู่ การฝึกหัดจะดำเนินต่อไปจนกว่าค่าความผิดพลาดจะลดลงต่ำกว่าค่าที่ยอมรับได้ (accept level) จึงจะหยุดฝึกหัด สุดท้ายค่าถ่วงน้ำหนักที่ได้จะเป็นเหมือนฟังก์ชันที่ใช้ในการแปลงข้อมูล

2) การเรียนรู้แบบไม่มีผู้สอน (Unsupervised learning)

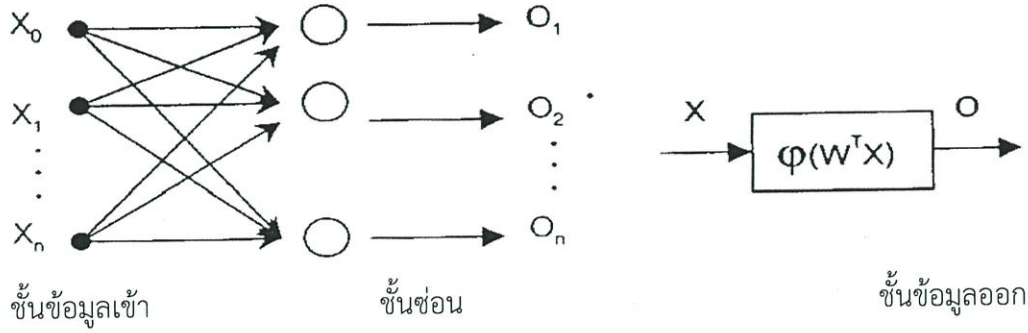
การเรียนรู้แบบไม่มีผู้สอนจะอาศัยชุดข้อมูลเข้าเพียงอย่างเดียวในการฝึกหัดโครงข่ายประสาทโดยไม่มีข้อมูลเป้าหมาย แต่จะใช้ข้อมูลออก (output data) จากโครงข่ายประสาทแทน เมื่อป้อนข้อมูลเข้าสู่โครงข่ายประสาท โครงข่ายประสาทจะคำนวณค่าความสัมพันธ์ที่มีอยู่ภายในกลุ่มข้อมูลเข้า โดยอาศัยค่าถ่วงน้ำหนักเป็นตัวแยกความแตกต่างของข้อมูลเข้าและนำไปเก็บไว้ในโหนดข้อมูลออกของโครงข่ายประสาท ซึ่งมีวัตถุประสงค์เพื่อใช้ในการจำแนกชุดข้อมูล (classification)

2.1.3.3 การเชื่อมโยงของโครงข่ายประสาทเทียม (Neural network linking)

เพื่อให้โครงข่ายประสาทสามารถเรียนรู้ได้อย่างมีประสิทธิภาพ จำเป็นต้องมีการเชื่อมโยงกันระหว่างเซลล์ประสาท โดยทั่วไปสามารถแบ่งการเชื่อมโยงของโครงข่ายได้ 2 ลักษณะคือ

1) โครงข่ายแบบไปข้างหน้า (Feedforward network)

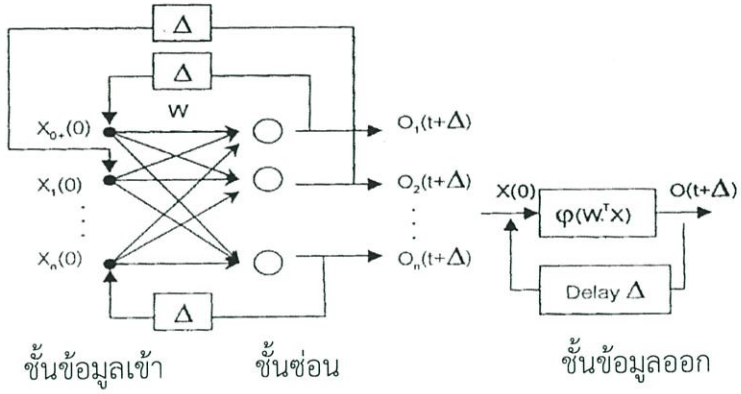
เป็นโครงข่ายที่การประมวลผลจะอาศัยชุดข้อมูลปัจจุบันและส่งค่าที่ประมวลผลได้ไปยังชั้นถัด ๆ ไป กล่าวคือ โครงข่ายชนิดนี้จะประกอบด้วยชั้นต่าง ๆ โดยชั้นแรกจะเป็นชั้นข้อมูลเข้า (input layer) และชั้นสุดท้ายเป็นชั้นข้อมูลออก (output layer) ส่วนระหว่างชั้นข้อมูลเข้ากับชั้นข้อมูลออกอาจจะมีหรือ ไม่มีชั้นซ่อน (hidden layer) อยู่ภายในก็ได้ ซึ่งขึ้นกับกฎการเรียนรู้ (learning rule) ที่ใช้ในการสอนโครงข่าย เช่น ถ้าเป็นโครงข่ายเพอร์เซปตรอนแบบหลายชั้น (multi-layer perceptron) จะมีชั้นซ่อนอยู่ระหว่างชั้นข้อมูลเข้ากับชั้นข้อมูลออก ซึ่งอาจมีมากกว่าหนึ่งชั้นได้ การเชื่อมต่อระหว่างชั้นของโครงข่ายแบบไปข้างหน้าจะมีค่าถ่วงน้ำหนัก (weight) เป็นตัวเชื่อมและสัญญาณนำเข้าที่เข้ามาจะถูกส่งไปตามทิศทางของลูกศรจนถึงชั้นข้อมูลออกโดยไม่มีการย้อนกลับ สามารถแสดงตัวแบบโครงข่ายแบบไปข้างหน้าได้ดังรูปที่ 2.4



รูปที่ 2.4 ลักษณะโครงข่ายประสาทเทียมแบบไปข้างหน้า

2) โครงข่ายแบบมีการย้อนกลับ (Feedback network)

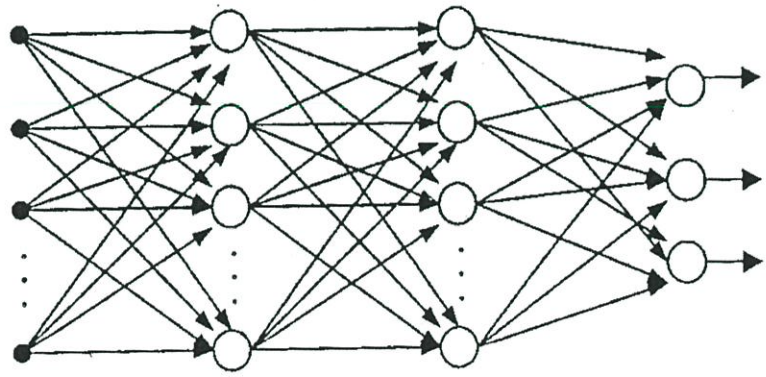
โครงข่ายชนิดนี้มีชื่อเรียกอีกชื่อหนึ่งว่า โครงข่ายหันกลับ (recurrent network) เป็นโครงข่ายที่จะอาศัยทั้งข้อมูลในปัจจุบันและข้อมูลที่มีการประวิงเวลามาใช้ในการประมวลผลของโครงข่ายประสาท สามารถแสดงตัวแบบโครงข่ายแบบมีการย้อนกลับได้ดังรูปที่ 2.5



รูปที่ 2.5 ลักษณะโครงข่ายประสาทเทียมแบบมีการย้อนกลับ

2.1.3.4 การแพร่แบบย้อนกลับ (Back-propagation)

การแพร่แบบย้อนกลับเป็นขั้นตอนที่ใช้สอนโครงข่ายประสาทแบบเพอร์เซปตรอนหลายชั้น (multi-layer perceptron) ซึ่งตัวแบบโครงข่ายประสาทมีการเชื่อมโยงกันเป็นโครงข่ายแบบเป็นชั้น ๆ โครงข่ายชนิดนี้มีการเชื่อมโยงกัน 3 ชั้น ประกอบด้วยชั้นข้อมูลเข้า (input layer) ถัดมาเป็นชั้นซ่อน (hidden layer) และชั้นสุดท้ายคือชั้นข้อมูลออก (output layer) สามารถแสดงโครงข่ายประสาทแบบเพอร์เซปตรอนหลายชั้นที่มีชั้นซ่อน 2 ชั้น ได้ดังรูปที่ 2.6



รูปที่ 2.6 ลักษณะโครงข่ายประสาทเทียมแบบเพอร์เซปตรอนหลายชั้น

ที่มาของชื่อการแพร่แบบย้อนกลับนั้นมาจากจุดที่ว่า วิธีการปรับค่าถ่วงน้ำหนักเพื่อให้ได้ค่าที่เหมาะสมนั้นจะใช้วิธีสอนว่าค่าเป้าหมาย (target) ของแต่ละข้อมูลเข้านั้นคืออะไร และใช้ค่า

ความผิดพลาด (error) ของข้อมูลออกมาใช้เป็นตัวชี้นำในการปรับค่าถ่วงน้ำหนัก ดังนั้นการแพร่แบบย้อนกลับ จึงเป็นกระบวนการเรียนรู้แบบมีผู้สอน แต่ปัญหาที่เกิดขึ้นคือไม่มีค่าเป้าหมายของสัญญาณที่ออกมาจากแต่ละเซลล์ประสาทในชั้นซ่อน ดังนั้นจึงต้องอาศัยการแพร่ความผิดพลาดจากชั้นข้อมูลออกกลับมายังชั้นซ่อนนั่นเอง

2.1.3.5 ปัจจัยที่ส่งผลต่อการเรียนรู้การแพร่แบบย้อนกลับ

1) การกำหนดค่าเริ่มต้นของค่าถ่วงน้ำหนัก

ก่อนที่จะทำการสอนโครงข่ายประสาทแบบหลายชั้น จำเป็นต้องกำหนดค่าเริ่มต้นให้กับค่าถ่วงน้ำหนักที่เชื่อมโยงระหว่างชั้นทุกชั้น โดยค่านี้จะเป็นเลขจำนวนจริงที่มีค่าน้อย ๆ ที่ได้มาจากการสุ่มค่าเริ่มต้น (randomness)

2) การกำหนดเกณฑ์การหยุดฝึกหัด

เกณฑ์ในการหยุดฝึกหัดนั้นขึ้นกับผู้ที่ทำการออกแบบโครงข่ายประสาทว่าต้องการที่จะให้โครงข่ายประสาทมีความแม่นยำเพียงใด โดยทั่วไปนิยมใช้ค่าดัชนีที่ชี้ถึงค่าความผิดพลาดของระบบได้ ในงานวิจัยส่วนใหญ่ใช้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error, MSE)

3) อัตราการเรียนรู้ (Learning rate, η)

อัตราการเรียนรู้เป็นค่าสัมประสิทธิ์ที่แสดงถึงการเรียนรู้ของโครงข่าย โดยทั่วไปค่าที่เหมาะสมจะอยู่ในช่วง 0.05 ถึง 0.5 ถ้าอัตราการเรียนรู้มีค่าสูง แสดงว่ากำหนดให้โครงข่ายมีการเปลี่ยนแปลงค่าถ่วงน้ำหนักที่มาก ในทางตรงกันข้ามถ้ามีอัตราการเรียนรู้ต่ำ แสดงว่ากำหนดให้โครงข่ายมีการเปลี่ยนแปลงค่าถ่วงน้ำหนักที่น้อย ซึ่งจำเป็นต้องใช้เวลาในการเรียนรู้ที่มากขึ้น แต่จะมีข้อดีคือโครงข่ายจะมีเสถียรภาพและไม่เกิดการแกว่ง (oscillation) ขณะที่ทำการเรียนรู้

4) ค่าคงที่โมเมนตัม (Momentum constant, α)

ค่าคงที่โมเมนตัมเป็นค่าสัมประสิทธิ์ที่ช่วยหน่วงไม่ให้การเปลี่ยนแปลงค่าถ่วงน้ำหนักนั้นมีค่ามากเกินไป เป็นการเพิ่มเสถียรภาพให้กับโครงข่ายประสาทได้อีกทางหนึ่ง ซึ่งค่าโมเมนตัมที่เหมาะสมจะมีค่าเข้าใกล้ 1.0 และควรจะกำหนดให้สอดคล้องกับอัตราการเรียนรู้ด้วย เช่น ถ้าอัตราการเรียนรู้สูงก็ควรที่จะมีค่าโมเมนตัมที่ต่ำ ทำให้การเปลี่ยนแปลงค่าถ่วงน้ำหนักนั้น ไม่มากจนเกินไป แต่ถ้าอัตราการเรียนรู้ต่ำก็ควรจะมีค่าโมเมนตัมที่สูง

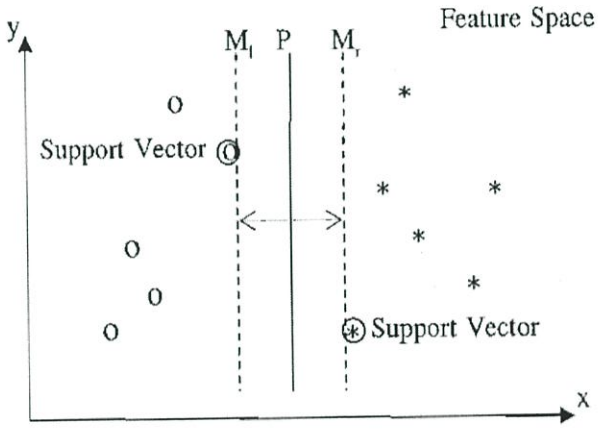
2.1.4 วิธีซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine Method)

เป้าหมายของวิธีซัพพอร์ทเวกเตอร์แมชชีนคือกระบวนการสอนเครื่องแบบมีผู้สอน (supervised learning) เพื่อให้สามารถสร้างตัวจำแนกข้อมูล (classifier) ที่มีความทั่วไป (generalize) สูง นั่นคือสามารถทำงานได้ดีกับตัวอย่างที่ไม่รู้จัก (unknown database) ด้วยกระบวนการปรับรูปแบบข้อมูลจากข้อมูลที่มีมิติต่ำ (low dimension dataset) บนพื้นที่ข้อมูลนำเข้า (input space) ให้อยู่ในรูปแบบของข้อมูลที่มีมิติสูง (high dimension dataset) บนพื้นที่ข้อมูลคุณลักษณะ (feature space) โดยใช้ฟังก์ชันในการปรับรูปแบบข้อมูลเรียกว่า **ฟังก์ชันเคอร์เนล (kernel function)** ซึ่งความสามารถดังกล่าวช่วยให้การสร้างตัวจำแนกข้อมูลด้วยสมการกำลังสอง (quadratic equation) บนพื้นที่ข้อมูลคุณลักษณะเป็นไปได้ง่ายขึ้นและมีความชัดเจนในการจำแนกกลุ่มมากยิ่งขึ้นด้วย นอกจากนี้ตัวจำแนกข้อมูลที่ดีควรมีโครงสร้างแบบเส้นตรง (linear classifier) และสามารถสร้างพื้นที่ระยะห่างระหว่างตัวจำแนกข้อมูลกับค่าที่ใกล้

ที่สุดของแต่ละกลุ่มข้อมูลได้มากที่สุดเพื่อประสิทธิภาพในการแยกแยะประเภทของชุดข้อมูลแต่ละประเภทออกจากกันอย่างชัดเจน ซึ่งเส้นที่เหมาะสมดังกล่าวเรียกว่า ระนาบแบ่งเขตข้อมูลที่เหมาะสม (optimal separating hyperplane)

2.1.4.1 แนวความคิดของซัพพอร์ทเวกเตอร์แมชชีน

ซัพพอร์ทเวกเตอร์แมชชีนเป็นสมการที่ใช้ในการจำแนกค่าคุณลักษณะของ 2 กลุ่ม ที่วางตัวอยู่ในพื้นที่คุณลักษณะ (feature space) ออกจากกันโดยจะสร้างเส้นแบ่ง (plane) ที่เป็นเส้นตรงขึ้นมาและเพื่อให้ทราบว่าเส้นตรงที่แบ่ง 2 กลุ่มออกจากกันนั้น เส้นตรงใดที่เป็นเส้นที่ดีที่สุด โดยเส้นตรงนั้นจะเพิ่มเส้นขอบ (margin) ออกไปทั้งสองข้าง โดยเส้นขอบที่เพิ่มนั้นจะขนานกับเส้นเดิมเสมอ เส้นขอบที่เพิ่มขึ้นมานี้จะขยายออกไปจนกว่าจะสัมผัสกับค่าของกลุ่มตัวอย่างที่ใกล้ที่สุด ดังรูปที่ 2.7



รูปที่ 2.7 การขยายตัวของเส้นขอบ

จากรูปที่ 2.7 เส้น M_l และ M_r คือเส้นขอบที่ขยายออกไปด้านซ้ายและขวาตามลำดับ และ P คือเส้นแบ่งข้อมูลทั้ง 2 กลุ่ม เมื่อเส้น M_l และ M_r ขยายออกจนไปสัมผัสค่าข้อมูลที่ใกล้ที่สุด ซึ่งข้อมูลที่อยู่บนเส้นขอบของทั้งสองฝั่งนั้นเรียกว่า **ซัพพอร์ทเวกเตอร์ (support vector)** จะวัดค่าระยะความห่างของเส้นขอบ โดยเส้น P จะเปลี่ยนความชันไปเรื่อย ๆ เพื่อที่จะหาความกว้างสูงสุดของเส้นขอบ

กระบวนการโดยรวมของซัพพอร์ทเวกเตอร์แมชชีนนั้นเป็นการหาค่าความชันของเส้น P ที่มีขนาดของเส้นขอบสูงสุด

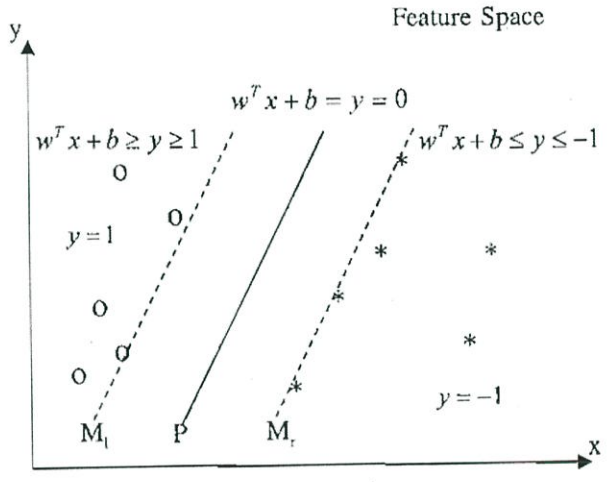
2.1.4.2 สมการพื้นฐานของซัพพอร์ทเวกเตอร์แมชชีน

ถ้านำแนวคิดของซัพพอร์ทเวกเตอร์แมชชีนที่กล่าวไปแล้วในข้อ 2.1.4.1 มาเขียนเป็นสมการเพื่อใช้ในการแก้ปัญหา โดยข้อมูลที่นำมาวางลงในพื้นที่คุณลักษณะนั้นเป็นกลุ่มข้อมูลที่อยู่ในรูปของเวกเตอร์

$$x = ((x_1, y_1),\ldots,(x_i, y_i)) \quad (2.5)$$

เมื่อ x คือ ชุดค่าคุณลักษณะ

ค่าคุณลักษณะที่วางตัวอยู่ในพื้นที่คุณลักษณะจะถูกแบ่งด้วยเส้นตรงดังรูปที่ 2.8 และเมื่อนำเส้นตรงมาแทนค่าด้วยสมการเส้นตรง $y = mx + b$ โดยมีการกำหนดกลุ่มของข้อมูลทั้งสองฝั่งเป็นเพียง 2 ค่า ที่ซึ่งแทนด้วยค่า y เพื่อให้ข้อมูลที่อยู่ในกลุ่มเดียวกันที่มาจากหลายค่ากลายเป็นค่าเดียว ดังสมการในรูปที่ 2.8



รูปที่ 2.8 เส้นขอบและเส้นแบ่งเมื่อแทนด้วยสมการเส้นตรง

จากรูปที่ 2.8 เส้นตรง M_l แทนด้วยสมการ $w^T x + b \geq y \geq 1$ ซึ่งข้อมูล y ที่มากกว่า 1 ก็จะถูกกำหนดค่าใหม่โดยให้ y เท่ากับ 1 และพจน์ w ก็คือค่าความชัน เช่นเดียวกันกับเส้นตรง M_r ที่ค่าของ y จะถูกกำหนดค่าใหม่เมื่อ y ที่น้อยกว่า −1 ให้เท่ากับ −1 ดังนั้นสมการที่เกิดขึ้นใหม่ จากสมการเส้นขอบ 2.6 และ 2.7 สามารถกำหนดได้ดังสมการที่ 2.8

เมื่อ $w^T x + b \geq y$ กำหนด y = 1 (2.6)

$w^T x + b \leq y$ กำหนด y = −1 (2.7)

$y(w^T x + b) - 1 \geq 0$ (2.8)

โดย y คือ ค่ากลุ่มข้อมูล (1, −1)

w คือ ค่าความชัน

x คือ ค่าคุณลักษณะ

b คือ ค่าคงที่ (ค่าตัดแกน y)

2.1.4.3 ค่าความกว้างเส้นขอบ (Margin)

การคำนวณความกว้างของเส้นขอบต้องทำการคำนวณพจน์ w ให้อยู่ในรูปปกติมาตรฐาน (normalization) โดยคำนวณจากสมการที่ 2.6 และ 2.7 เมื่อแทนค่า y ลงไปแล้ว

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

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

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

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

$$= \frac{2}{\|w\|} \quad (2.9)$$

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i \quad (2.10)$$

โดยที่ M คือ ความกว้างของเส้นขอบ

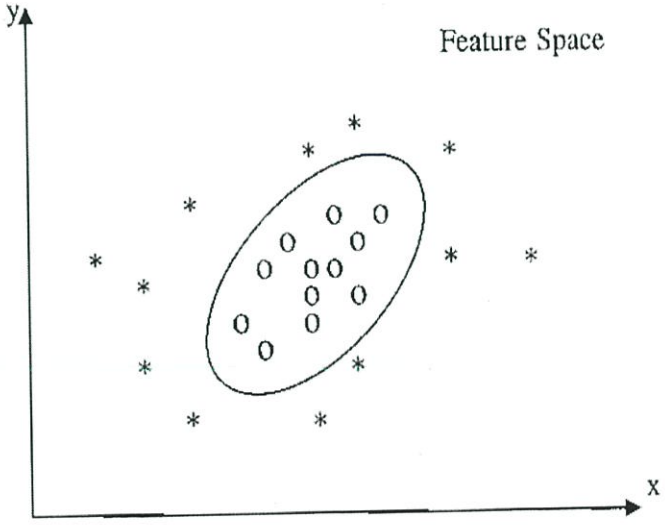
α คือ สัมประสิทธิ์คงที่

เมื่อนำค่า w ไปใส่ในสมการที่ 2.8 ซึ่งเป็นสมการในการหาเส้นแบ่ง จะได้

$$y_i\left(\sum_{i=1}^{N} \alpha_i y_i \left(x_i^T, x_j\right) + b\right) - 1 \geq 0 \quad (2.11)$$

2.1.4.4 เคอร์เนล (Kernel)

ในความเป็นจริงนั้นข้อมูล 2 กลุ่ม ไม่ได้วางตัวในพื้นที่คุณลักษณะ และไม่สามารถแบ่งได้โดยเส้นตรง แต่ข้อมูลอาจจะจับกลุ่มกันในตำแหน่งต่าง ๆ ดังนั้นจึงเป็นปัญหาทำให้ไม่สามารถที่จะใช้สมการซัพพอร์ทเวกเตอร์แมชชีนแบบเชิงเส้นได้ ดังนั้นจะต้องมีเครื่องมือมาช่วยให้ข้อมูลเหล่านั้นเรียงตัวใหม่ในพื้นที่ เรียกว่า **พื้นที่หลายมิติ (higher dimensional space)**



รูปที่ 2.9 รูปแบบการวางตัวที่ไม่สามารถแบ่งด้วยเส้นตรงได้

ในเคอร์เนลนั้นคือการคูณกันของชุดเวกเตอร์ของ x ใด ๆ

$$K\left(x_i, x_j\right) = x_i^T x_j \quad (2.12)$$

เคอร์เนลที่นิยมใช้มีอยู่ 3 ชนิด คือ

1) โพลิโนเมียล (Polynomial)

$$K\left(x_i, x_j\right) = \left(< x_i, x_j > +1\right)^d \quad (2.13)$$

เมื่อ d คือ ค่าเลขยกกำลัง

2) ฟังก์ชันเบสิสเรเดียล (Radial Basis Function : RBF)

$$K(x_i, x_j) = \exp\left(-\frac{x_i - x_j^2}{2\sigma^2}\right) \quad (2.14)$$

เมื่อ σ คือ ค่าพารามิเตอร์

3) ซิกมอยด์ (Sigmoid)

$$K(x_i, x_j) = \tanh(k < x_i, x_j > +\mu) \quad (2.15)$$

เมื่อ k, μ คือค่าพารามิเตอร์

ดังนั้นจากสมการของเคอร์เนลนั้นสามารถที่จะแทนลงไปในตำแหน่งของ x_i^T, x_j ในสมการที่ 2.11 จึงเขียนเป็นสมการใหม่ดังนี้

$$y_i \sum_{i=1}^{N}(\alpha_i y_i K(x_i, x_j) + b) - 1 \geq 0 \quad (2.16)$$

สมการที่ 2.16 เป็นสมการที่ใช้ในขั้นตอนที่จะเรียนรู้ว่าจะวางตำแหน่งเส้นแบ่งไว้ที่ตำแหน่งใด โดยทำงานร่วมกับเคอร์เนล เพื่อแปลงให้ข้อมูลที่ยากต่อการแบ่งแบบเชิงเส้นสามารถแบ่งได้เมื่อทำให้เป็นข้อมูลแบบหลายมิติ (higher dimension) ดังนั้นจึงมีอีกสมการหนึ่งที่ใช้ค่า w และ b เดิมมาจัดตำแหน่งของข้อมูลเพื่อที่ให้ทราบว่าข้อมูลนั้นเป็นกลุ่มใด กำหนดได้ดังสมการที่ 2.17

$$f(x) = \text{sgn}\left(\sum_{i=1}^{\infty}(\alpha_i y_i K(x_i, x_j) + b)\right) \quad (2.17)$$

เมื่อ $f(x)$ คือค่า y หาในรูปของ x

2.1.5 วิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม (Binary Logistic Regression Method) เป็นการวิเคราะห์การถดถอยแบบหนึ่งโดยที่ตัวแปรตามเป็นตัวแปรเชิงคุณภาพที่มีค่าได้เพียง 2 ค่า (dichotomous or binary variable) ส่วนตัวแปรอิสระอาจจะเป็นตัวแปรเชิงปริมาณหรือเชิงคุณภาพ หรืออาจจะมีทั้งตัวแปรเชิงปริมาณและตัวแปรเชิงคุณภาพก็ได้ (กัลยา วานิชย์บัญชา, 2552)

2.1.6 วิธีนาอีฟ เบย์ (Naïve Bayes Method) คืออัลกอริทึมที่ใช้หลักการของความน่าจะเป็นในการคัดกรองแต่ละคำตอบ (Class) โดยมีคำตอบ 2 คำตอบ (สายชล สินสมบูรณ์ทอง, 2558)

2.2 รายงานวิจัยที่เกี่ยวข้อง

ณัฐวรรณ รัตนากรกุล (2545) ได้ทำวิจัยเรื่องระบบจำแนกประเภทโครงสร้างของโปรตีนโดยใช้เทคนิค Data Mining มีจุดประสงค์เพื่อพัฒนาระบบจัดแบ่งประเภทโครงสร้างของโปรตีนเพื่อรองรับการค้นพบชนิดของโปรตีนใหม่ ๆ งานวิจัยนี้ได้ใช้เทคนิคการทำเหมืองข้อมูล 2 เทคนิค คือ การจัดแบ่งประเภทข้อมูล (Data Classification) และการค้นหากฎความสัมพันธ์ (Association Rule Discovery) มาประยุกต์ใช้ในการสร้างระบบจัดแบ่งประเภทโครงสร้างของ

โปรตีน ข้อมูลที่นำมาเรียนรู้เพื่อสร้างระบบคือข้อมูลเกี่ยวกับภาพแบบลำดับกรดอะมิโนที่มีความสอดคล้องในกลุ่มโปรตีน ข้อมูลเกี่ยวกับฟังก์ชันการทำงานของโปรตีนและแหล่งที่มาของโปรตีน ผลที่ได้สามารถสร้างกฎความสัมพันธ์ได้ระหว่างประเภทข้อมูล

กิตติพล วิแสง และคณะ (2552) ได้ศึกษาเรื่องการวิเคราะห์ปัจจัยเสี่ยงของโรคเบาหวาน พบว่าปัจจัยเสี่ยงของโรคเบาหวานมีความสำคัญอย่างยิ่งในการนำมาเป็นเครื่องมือประเมินการเกิดโรคเบาหวาน แทนการตรวจเลือด บทความนี้ศึกษาและวิเคราะห์ปัจจัยเสี่ยงและลำดับของปัจจัยเสี่ยงในการเกิดโรคเบาหวาน โดยตัวแบบที่ใช้ศึกษาคือตัวแบบ Back-propagation Neural Networks, Radial Basis Function Network และตัวแบบ Naïve Bayes ผลการวินิจฉัยทางการแพทย์ได้ถูกนำมาอ้างอิงเพื่อวัดความถูกต้องของตัวแบบ ผลการศึกษาแสดงให้เห็นว่าตัวแบบ Back-propagation Network ให้ผลลัพธ์ที่มีความถูกต้องมากที่สุด ปัจจัยเสี่ยงและลำดับของปัจจัยเสี่ยงสามารถตัดสินได้จากตัวแบบที่ศึกษา ซึ่งข้อมูลเหล่านี้สามารถนำไปสร้างเป็นแบบประเมินความเสี่ยงโรคเบาหวานโดยไม่ต้องอาศัยการตรวจเลือดได้ รวมทั้งเป็นเครื่องมือในการประเมินตนเอง และสามารถพัฒนาเป็นระบบผู้เชี่ยวชาญสำหรับการวินิจฉัยโรคเบาหวานต่อไปได้

ภัทรพงศ์ พงศ์ภัทรกานต์ (2552) ได้นำเสนอการเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูลของแบบจำลอง C5.0, CART, SVM และ SVM ร่วมกับ C5.0 ภายใต้หลักการทำงานของเหมืองข้อมูลโดยใช้ชุดข้อมูลจำนวน 9 ชุด ทำการเลือกชุดข้อมูลที่เป็นไบนารีคลาสและมีคุณลักษณะ (attribute) กับจำนวนตัวอย่างที่หลากหลายรูปแบบ SVM จะทำการคัดแยกชุดข้อมูลออกเป็นไบนารีคลาสได้ดีมาเปรียบเทียบประสิทธิภาพในการจำแนกกลุ่มข้อมูลโดยใช้ C5.0, CART, SVM และ SVM ผสมผสานกับ C5.0 โดยทำการทดลองวัดประสิทธิภาพความถูกต้องเปรียบเทียบกัน ซึ่งผลการทดลองพบว่าแบบจำลอง SVM ผสมผสานกับ C5.0 ที่ผู้วิจัยนำเสนอมีประสิทธิภาพสูงที่สุดทุกชุดข้อมูลจากจำนวน 9 ชุด ที่ได้ทำการทดลอง ซึ่งสรุปผลได้ว่าการใช้ตัวแบบผสมผสานกันสามารถจำแนกประเภทข้อมูลประเภทไบนารีคลาสได้อย่างมีประสิทธิภาพสูงกว่าการใช้ตัวแบบ C5.0, CART และ SVM เพียงอย่างเดียว

ภัทราวุฒิ แสงศิริ (2553) ได้ทำวิจัยเรื่องการคัดแยกประเภทของมะเร็งเม็ดเลือดขาวโดยใช้วิธีการจัดอันดับร่วมกับเทคนิคซัพพอร์ตเวกเตอร์แมชชีน โดยการค้นหากลุ่มย่อยของยีนที่มีอำนาจจำแนก เป็นปัญหาที่สำคัญสำหรับงานวิจัยทางด้านชีววิทยา เนื่องจากมีจำนวนยีนเพิ่มขึ้นเป็นจำนวนมาก ดังนั้นเทคนิคการลดมิติของข้อมูลจึงเป็นประโยชน์ในการช่วยค้นหากลุ่มย่อยของยีน สาเหตุเนื่องจากเมื่อข้อมูลมีจำนวนมิติหรือตัวแปรมาก ทำให้ข้อมูลเกิดการกระจาย (data sparse) และทำให้เกิดปัญหามิติข้อมูล (Curse of Dimensionality) งานวิจัยนี้จะนำเอาข้อมูลยีนของโรคมะเร็งเม็ดเลือดขาวแบบเฉียบพลัน (Acute Leukemia) ซึ่งมีจำนวนมิติของข้อมูล 7,129 มิติ แบ่งออกเป็น 2 กลุ่ม คือ ALL และ AML มาทำการทดลองและเพื่อเปรียบเทียบประสิทธิภาพของการลดมิติข้อมูล ระหว่างวิธี Correction Based Feature Selection, Gain Ratio และ Information Gain โดยนำผลลัพธ์ที่ได้จากการลดมิติมาเป็นข้อมูลอินพุตของซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) เพื่อคัดแยกประเภทของโรคมะเร็ง ซึ่งผลการทดลองแสดงให้เห็นว่าการลดข้อมูลโดยวิธี Gain Ratio และ Information Gain มีความเหมาะสม

นิเวศ จิระวิชิตชัย (2553) ได้ทำการศึกษาเรื่องการค้นหาเทคนิคการทำเหมืองข้อมูลเพื่อสร้างโมเดลการวิเคราะห์โรคอัตโนมัติ โดยงานวิจัยนี้มีวัตถุประสงค์เพื่อค้นหาเทคนิคด้านเหมืองข้อมูล

เพื่อสร้างโมเดลการวิเคราะห์โรคอัตโนมัติทดสอบประสิทธิภาพในการจำแนก (Classification) สำหรับข้อมูลทางการแพทย์โดยทดลองกับ 7 อัลกอริทึม ซึ่งประกอบด้วย Naïve Bayes, Multilayer Perceptron, Radial Basis Function Network, Support Vector Machine, K-Nearest Neighbor, Decision Tree, Ripper ทำการศึกษาเปรียบเทียบวิธีลดคุณลักษณะที่เหมาะสมด้วยวิธี Correction-based Feature Subset Selection (CFS) และวิธี Feature selection method based on correction measure and relevance & redundancy analysis (FCBF) รวมถึงทดสอบอัลกอริทึมประเภท Single learning และ Multiple learning และทำการเพิ่มประสิทธิภาพการจำแนกด้วยวิธี Bagging และ Boosting ผลการวิจัยพบว่าทุกโมเดลที่สร้างขึ้นมีประสิทธิภาพในการจำแนกประเภทของโรคในระดับ 80% ขึ้นไป เมื่อไม่ลดคุณลักษณะ และเมื่อเรียงค่าความถูกต้อง (Accuracy) แยกตามประเภทของข้อมูลพบว่ากลุ่มข้อมูล Hypothyroid การสร้างโมเดลด้วยอัลกอริทึม Decision Tree ให้ประสิทธิภาพดีที่สุด 99.57% กลุ่มข้อมูล Leukemia การสร้างโมเดลด้วยอัลกอริทึม Naïve Bayes กับ Support Vector Machine ให้ประสิทธิภาพดีที่สุด 98.61% กลุ่มข้อมูล Breast-w การสร้างโมเดลด้วยอัลกอริทึม Support Vector Machine ให้ประสิทธิภาพดีที่สุด 96.66% กลุ่มข้อมูล Lymphography การสร้างโมเดลด้วยอัลกอริทึม Support Vector Machine ให้ประสิทธิภาพดีที่สุด 86.48% กลุ่มข้อมูล Hepatitis การสร้างโมเดลด้วยอัลกอริทึม Radial Basis Function กับ K-Nearest Neighbor ให้ประสิทธิภาพดีที่สุด 85.80% กลุ่มข้อมูล Heart-c การสร้างโมเดลด้วยอัลกอริทึม Support Vector Machine ให้ประสิทธิภาพดีที่สุด 84.15% กลุ่มข้อมูล Heart-statlog การสร้างโมเดลด้วยอัลกอริทึม Radial Basis Function กับ Support Vector Machine ให้ประสิทธิภาพดีที่สุด 84.07% ตามลำดับ ซึ่งการสร้างโมเดลดังกล่าวมีคุณภาพในระดับที่ยอมรับได้และสามารถนำไปพัฒนาเป็นซอฟแวร์ในการวินิจฉัยโรคอัตโนมัติได้ การเพิ่มประสิทธิภาพด้วยวิธี Multiple Learning ด้วยอัลกอริทึม Bagging และ Boosting ส่งผลให้ค่าความถูกต้องเพิ่มขึ้นเฉพาะบางกลุ่มข้อมูลเท่านั้น โดยมีข้อสังเกตว่าสัดส่วนของกลุ่มตัวอย่างแต่ละคลาสจะต้องมีปริมาณใกล้เคียงกันหรือเท่ากัน กรณีที่การกระจายของกลุ่มตัวอย่างแต่ละคลาสมีสัดส่วนแตกต่างกันมากส่งผลให้เทคนิค Bagging และ Boosting ไม่ช่วยเพิ่มประสิทธิภาพในการจำแนกข้อมูล การลดคุณลักษณะด้วยวิธี Correction-based Feature Subset Selection (CFS) และวิธี Feature selection method based on correction measure and relevance & redundancy analysis (FCBF) ส่งผลให้ประสิทธิภาพความถูกต้อง (Accuracy) ในการจำแนกประเภทของโรคใกล้เคียงกับการไม่ลดคุณลักษณะ แต่การลดมิติของข้อมูลดังกล่าวทำให้ประหยัดทรัพยากรของระบบคอมพิวเตอร์และระยะเวลาในการเรียนรู้เพื่อสร้างโมเดลได้เป็นอย่างดี

วาทินี นุ้ยเพียร และคณะ (2553) ได้ทำการศึกษาเพื่อเปรียบเทียบวิธีการจำแนกข้อมูลโดยเลือกใช้อัลกอริทึมโครงข่ายประสาทแบบมัลติเลเยอร์เพอร์เซ็ปตรอน ซัพพอร์ตเวกเตอร์แมชชีน นาอีฟเบย์และความใกล้เคียงกันมากที่สุดเพื่อประเมินประสิทธิภาพค่าความถูกต้อง (accuracy) ค่าความแม่นยำ (precision) ค่าความระลึก (recall) และค่าความถ่วงดุล (F-measure) ใช้ข้อมูลจาก UCI ประกอบด้วย Ozone Days และ Adult เลือกกลุ่มข้อมูลโดยมีจำนวนคำตอบ (class) เท่ากันในข้อมูลแต่ละชุด เป็นการทดลองแบบมีการเรียนรู้ จากผลการวิจัยอัลกอริทึมที่ดีที่สุดของข้อมูล Ozone Days คือซัพพอร์ตเวกเตอร์แมชชีน ฟังก์ชันเคอร์เนลแบบ Rbf มีค่าความถูกต้อง 94.83% ค่าความแม่นยำ 96% ค่าความระลึก 96% และค่าความถ่วงดุล 96%

ส่วนข้อมูล Adult คือซัพพอร์ตเวกเตอร์แมชชีน ฟังก์ชันเคอร์เนลแบบโพลิโนเมียล มีค่าความถูกต้อง 79.66% ค่าความแม่นยำ 80% ค่าความระลึก 80% และค่าความถ่วงดุล 80% อัลกอริทึมที่สามารถเลือกใช้ได้ดีคือ ซัพพอร์ตเวกเตอร์แมชชีน สามารถใช้กับลักษณะข้อมูลที่เป็นตัวเลขหรือข้อมูลเชิงกลุ่มแบบข้อความ ซึ่งเทคนิคเหล่านี้สามารถประยุกต์ใช้กับการสร้างเทคโนโลยีการจัดเก็บและนำเสนอเนื้อหาแบบมีโครงสร้าง โดยสามารถวิเคราะห์จำแนกหรือจัดแบ่งข้อมูลที่มีความสัมพันธ์กับข้อมูลอื่น ๆ แบบเชิงความหมายได้ต่อไป

ชาญชัย พรมโคตร และคณะ (2554) ศึกษาการวิเคราะห์ข้อมูลความเสี่ยงของการเป็นโรคหอบหืดด้วยเทคนิคโครงข่ายประสาทเทียมแบบหลายชั้นร่วมกับขั้นตอนวิธีการเรียนรู้แบบแพร่ย้อนกลับ พบว่าโรคหอบหืดเป็นอาการป่วยที่ร้ายแรงอาจทำให้เสียชีวิตถ้าผู้ป่วยไม่ปฏิบัติตัวให้ถูกต้องจนอาการรุนแรง การตรวจผู้ป่วย ทำโดยผู้เชี่ยวชาญประกอบกับผลทางห้องปฏิบัติการในรายที่มีอาการแบบเฉียบพลันจะส่งผลให้การรักษาผู้ป่วยไม่ทันท่วงที จนเกิดอาการรุนแรงและอาการแทรกซ้อน แต่ถ้าผู้ป่วยทราบว่าตนเองเป็นโรคหอบหืดตั้งแต่ระยะแรกของโรค และได้รับการตรวจรักษา พร้อมทั้งปฏิบัติตนอย่างเหมาะสม จะทำให้ลดความเสียงของการเสียชีวิตและภาวะแทรกซ้อนได้ การจำแนกข้อมูลผู้ป่วยซึ่งเป็น ข้อมูลอาการ เช่น ไอ แน่นหน้าอก หายใจหอบเหนื่อย ด้วยเทคนิคโครงข่ายประสาทเทียมแบบหลายชั้น (Multi-Layer Perceptron) ร่วมกับขั้นตอนวิธีการเรียนรู้แบบแพร่ย้อนกลับ (Back Propagation Learning Algorithm) จึงถูกนำเสนอเพื่อวิเคราะห์ความเสียงของการเกิดหอบหืด โดยวิธีการนี้มีความถูกต้องมากกว่าร้อยละ 90 และสามารถนำไปใช้ประกอบการตัดสินใจของผู้ป่วยเพื่อเข้ารับการรักษาตั้งแต่เริ่มต้นได้

สุมนพิชญ์ พลศรี และอัจฉรา มหาวีรวัฒน์ (2554) ศึกษาการวิเคราะห์ความเสี่ยงของการเป็นวัณโรคด้วยเทคนิคโครงข่ายประสาทเทียม พบว่าวัณโรคเป็นโรคติดต่อเรื้อรังรักษาให้หายได้ แต่ถ้าไม่กินยาให้ครบตามกำหนดอาจทำให้เชื้อดื้อยายากแก่การรักษา แต่ถ้าหากผู้ป่วยทราบว่าตนเองเป็นวัณโรคในระยะแรก ๆ จะช่วยลดความรุนแรงของโรคก่อนที่จะเป็นอันตรายถึงแก่ชีวิตหรือสามารถป้องกันไม่ให้โรคติดต่อไปสู่บุคคลอื่นได้ วิธีการวิเคราะห์ความเสี่ยงของการเกิด วัณโรคจึงถูกนำเสนอ โดยประยุกต์ใช้เทคนิคโครงข่ายประสาทเทียมแบบหลายชั้น (Multi-Layer Perceptron) ร่วมกับขั้นตอนวิธีการเรียนรู้แบบแพร่ย้อนกลับ (Back Propagation Learning Algorithm) เพื่อจำแนกข้อมูลผู้ป่วย ซึ่งได้แก่ ข้อมูลอาการ ประวัติการเป็นโรคของสมาชิกในครอบครัวและปัจจัยเสี่ยงที่มีผลต่อการเป็นวัณโรค จากผลการวิเคราะห์พบว่าวิธีการนี้มีความถูกต้องแม่นยำและสามารถนำไปใช้ประกอบการตัดสินใจของผู้ป่วยเพื่อเข้ารับการรักษาตั้งแต่ระยะแรกของโรคได้

เดช ธรรมศิริ และพยุง มีสัจ (2554) ได้ทำการศึกษาเรื่องการเรียนรู้แบบรวมกลุ่มด้วยโครงข่ายประสาทเทียมเอดาบูทสำหรับการจำแนกข้อมูล พบว่าในปัจจุบันการเพิ่มความแม่นยำสำหรับการจำแนกข้อมูลถือเป็นเรื่องสำคัญ ดังนั้นงานวิจัยครั้งนี้เป็นการนำเสนอวิธีการจำแนกข้อมูลด้วยวิธีร่วมกันตัดสินใจโดยใช้เทคนิคโครงข่ายประสาทเทียมร่วมกับการรวมกลุ่มตัดสินใจโดยใช้เทคนิคเอดาบูทเพื่อให้ความถูกต้องในการจำแนกข้อมูลที่สูงขึ้น ทดสอบบนฐานข้อมูลโรคเบาหวานได้แก่ Diabetes Data จาก UCI การทดสอบความถูกต้องในการจำแนกข้อมูล Diabetes Data พบว่าเทคนิคการร่วมกันตัดสินใจจากหลายโมเดลที่ผ่านการรวมกลุ่มตัดสินใจด้วยเทคนิคเอดาบูทมีผลลัพธ์ที่ดีกว่าการใช้เทคนิคแบบโมเดลเดี่ยว โดยผลการวิจัยพบว่าสำหรับข้อมูล Diabetes Data ให้ผลลัพธ์ความถูกต้องสูงสุดที่ 75.02% ในขณะที่โมเดลเดี่ยวให้ความแม่นยำที่น้อยกว่า

นงเยาว์ ในอรุณ และพรรณี สิทธิเดช (2555) ได้ศึกษาการใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีนและโครงข่ายประสาทเทียมแบบแพร่กลับสำหรับการจำแนกผู้ป่วยโรคหัวใจขาดเลือดและโรคหัวใจรูปแบบอื่น โดยงานวิจัยนี้มีวัตถุประสงค์เพื่อจำแนกผู้ป่วยโรคหัวใจขาดเลือดและโรคหัวใจรูปแบบอื่นออกจากกันให้ชัดเจน เนื่องจากสองโรคนี้มีลักษณะอาการที่คล้ายกัน ทำให้แพทย์วินิจฉัยโรคได้ยาก ในปัจจุบันโรคหัวใจขาดเลือดเป็นโรคที่อันตรายและทำให้ผู้ป่วยทั่วโลกเสียชีวิตเป็นจำนวนมากและมีแนวโน้มเพิ่มขึ้นอย่างต่อเนื่อง ถ้าแพทย์สามารถตรวจพบโรคได้ก่อนจะช่วยลดความเสี่ยงของการเสียชีวิตของผู้ป่วยได้ วิธีการดำเนินการวิจัยได้รวบรวมข้อมูลผู้ป่วยจำนวน 2,500 ระเบียน มาทำการให้เป็นมาตรฐานเดียวกัน (Min-max normalization) และใช้อัลกอริทึม K-means เพื่อขจัดข้อมูลที่ไม่ชัดเจนออกไป ซึ่งได้ข้อมูลที่ถูกต้องสำหรับนำมาใช้ทดลองจำนวน 1866 ระเบียน สุ่มแบ่งข้อมูลออกเป็นสองกลุ่ม กลุ่มละ 933 ระเบียน เพื่อนำมาสร้างโมเดลและใช้ทดสอบโมเดล ใช้ข้อมูลชุดแรกเพื่อสร้างโมเดลด้วยเทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support vector machine : SVM) และเทคนิคโครงข่ายประสาทเทียมแบบแพร่กลับ (Back propagation neural network : BPNN) เมื่อเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลทั้งสองวิธีพบว่าเทคนิค SVM ได้ค่าความถูกต้อง (96.46%) มากกว่าเทคนิค BPNN (88.21%) สรุปได้ว่าเทคนิค SVM เป็นเทคนิคที่เหมาะสมในการจำแนกผู้ป่วยโรคหัวใจขาดเลือดออกจากโรคหัวใจรูปแบบอื่น

เพิ่มพันธุ์ ธรรมสโรช และพรภัทร ธรรมสโรช (2555) ได้ศึกษาการทำนายการเกิดภาวะเลือดออกในสมองจากการใช้ยาละลายลิ่มเลือดในผู้ป่วยโรคหลอดเลือดสองตีบและอุดตันโดยใช้แบบจำลองโครงข่ายประสาทเทียมหลายแบบ พบว่าโครงข่ายประสาทเทียมได้รับการนำมาใช้มากขึ้นในการวินิจฉัยและการทำนายปัจจัยการตายและความเสี่ยงในโรคหลอดเลือดสองตีบแต่ละแบบจำลองอาจมีความแม่นยำ ความไวและความจำเพาะที่แตกต่างกันในการประมวลผลข้อมูลทางคลินิกที่เหมือนกัน ดังนั้นการใช้เพียงหนึ่งรูปแบบของโครงข่ายประสาทเทียมอาจทำให้เกิดความคลาดเคลื่อนได้ การศึกษานี้มีวัตถุประสงค์เพื่อทำนายภาวะเลือดออกในสมองภายหลังการใช้ยาละลายลิ่มเลือดในโรคหลอดเลือดสมองตีบเฉียบพลันโดยใช้แบบจำลองโครงข่ายประสาทเทียมหลายแบบ คือ แบบจำลอง radial basis function (อาร์บีเอฟ) multilayer perceptron (เอ็มแอลพี) probabilistic neural network (พีเอ็นเอ็น) และ support vector machine (เอสวีเอ็ม) ถูกสร้างขึ้นเพื่อวิเคราะห์ 194 ชุดข้อมูล ที่มีตัวแปร 29 ตัวแปร การทำนายความสำคัญของตัวแปรถูกคำนวณโดยใช้การวิเคราะห์ความไว การเปรียบเทียบระหว่างแบบจำลองพบว่าไม่มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติในผลการทำนายของอาร์บีเอฟ เอ็มแอลพี พีเอ็นเอ็น และเอสวีเอ็มแบบจำลองพีเอ็นเอ็นมีประสิทธิภาพที่ดีกว่าแบบจำลองเอสวีเอ็มอย่างมีนัยสำคัญ เมื่อพิจารณาคะแนนความสำคัญที่ต่ำสุดที่ 50 ร่วมกับค่าพื้นที่ใต้กราฟ > 0.50 พบว่ามีสามแบบจำลองที่แสดงว่าประเภทของโรคหลอดเลือดสองเป็นตัวแปรทำนายที่สำคัญสำหรับภาวะเลือดออกในสมอง ตัวแปรอื่น ๆ ที่มีศักยภาพในการทำนายได้แก่ ตำแหน่งของรอยโรค ค่าเวลาโปรธรอมบิน ระดับคลอเรสเตอรอล ระดับไขมันแอลดีแอล ระดับความดันระยะหัวใจคลายตัว ค่าอัตราส่วนไอเอ็นอาร์ของการแข็งตัวของเลือดและเอ็กซ์เรย์คอมพิวเตอร์สมอง สรุปได้ว่าถึงแม้ว่าแบบจำลองโครงข่ายประสาทเทียมจะมีประสิทธิภาพการทำงานที่คล้ายกัน แต่ผลการจัดหมวดหมู่ไม่เหมือนกันทั้งหมด บ่งบอกถึงข้อดีของการใช้แบบจำลองการจัดหมวดหมู่หลายแบบมากกว่าแบบเดียว ผลการศึกษาเชิงสถิติก่อนหน้านี้สนับสนุนผลการทำนายนี้ การศึกษานี้แสดงให้เห็นว่าโครงข่ายประสาทเทียมสามารถนำไปใช้โดยทั่วไป

เสกสันติ จันทะมงคล และคณะ (2555) ได้ทำการศึกษาเรื่องระบบช่วยเหลืออัจฉริยะเพื่อการวางแผนการรักษาโรคเรื้อรังโดยใช้เหมืองข้อมูล พบว่าโรคเบาหวานเป็นโรคเรื้อรังที่ต้องได้รับการดูแลรักษาอย่างต่อเนื่องตลอดชีวิตของผู้ป่วย ในปัจจุบันผู้ป่วยโรคเบาหวานมีจำนวนเพิ่มมากยิ่งขึ้น การรักษาผู้ป่วยโรคเบาหวานรายใหม่นั้นแพทย์จำเป็นต้องใช้ประวัติการรักษาผู้ป่วยรายอื่นที่เป็นกลุ่มอาการเดียวกันที่รักษาไปแล้ว เพื่อวางแผนการรักษาที่เหมาะสมกับผู้ป่วยรายใหม่ แพทย์จึงมีความจำเป็นที่จะต้องมีการวินิจฉัยสั่งการดูแลรักษาอย่างรวบรัดและบางครั้งการรักษาอาจไม่ครบถ้วนในกลุ่มโรคนั้น ทำให้ผลการรักษาผู้ป่วยไม่เป็นไปตามที่คาดไว้และประสบปัญหาการวางแผนการรักษาของแพทย์ งานวิจัยนี้ได้วิเคราะห์ข้อมูลประวัติกลุ่มผู้ป่วยโรคเรื้อรังเพื่อให้ได้รูปแบบ (Model) ของแผนการรักษาที่เหมาะสมกับแต่ละลักษณะของผู้ป่วยโรคเรื้อรัง โดยใช้การทำเหมืองข้อมูล ผลที่ได้รับคือช่วยให้การวางแผนการรักษาโรคเรื้อรังของแพทย์มีประสิทธิผลมากขึ้น ทั้งด้านความแม่นยำ เวลาในการวางแผนและการประเมินผลจะช่วยให้ประสิทธิภาพการให้บริการของโรงพยาบาลสูงขึ้นและมีประสิทธิผลมากขึ้น จากผลการทดลองนี้สรุปได้ว่าตัวแบบที่สร้างให้ค่าพยากรณ์โดยพิจารณาจากค่าทางสถิติที่ได้จากการวิเคราะห์และทดสอบค่าความถูกต้อง Correctly Classified Instances และค่าความคลาดเคลื่อน Root Mean Square Error (RMSE) พิจารณาตัวแบบในการเรียนรู้จากอัลกอริทึมต้นไม้ตัดสินใจในข้อมูลชุดกลุ่มผู้ป่วยโรคเบาหวานได้ผลได้นี้ ค่าความถูกต้องสูงสุดที่ 83.05 เปอร์เซ็นต์ และค่าเฉลี่ยความคลาดเคลื่อนต่ำสุดที่ 16.95 เปอร์เซ็นต์ Root Mean Square Error (RMSE) เป็นดัชนีที่ใช้วัดความคลาดเคลื่อนระหว่างค่าจริงและค่าที่พยากรณ์ได้ พบว่ามีค่าความคลาดเคลื่อนเพียง 14.37 เปอร์เซ็นต์

วีระยุทธ มายุศิริ และคณะ (2556) ได้ทำการศึกษาการพัฒนาแบบจำลองเพื่อการพยากรณ์การรักษาซ้ำของผู้ป่วยโรคจิตเภทโดยเทคนิคเหมืองข้อมูล พบว่าจิตเภท (Schizophrenia) เป็นโรคทางจิตเวชที่พบมากกว่าการเจ็บป่วยด้วยโรคจิตประเภทอื่น ๆ และพบว่าผู้ป่วยเหล่านี้ส่วนใหญ่มักกลับมาเป็นซ้ำ (relapse) ต้องกลับเข้ารักษาในโรงพยาบาลเป็นระยะ ทำให้ผู้ดูและรัฐบาลต้องสูญเสียงบประมาณในการดูแลรักษาเป็นจำนวนมาก งานวิจัยนี้มีวัตถุประสงค์พัฒนาแบบจำลองในการพยากรณ์ระยะเวลาการรักษาซ้ำของผู้ป่วยโรคจิตเภทโดยเทคนิคเหมืองข้อมูล จากฐานข้อมูลโรงพยาบาลพระศรีมหาโพธิ์ จังหวัดอุบลราชธานี ปี พ.ศ. 2550 ถึงปี พ.ศ. 2555 จำนวน 2831 เรคคอร์ด โดยการแบ่งข้อมูลออกเป็น 2 คลาส คือ คลาส 0 เป็นกลุ่มของผู้ป่วยมารับการรักษาซ้ำใน 1 ถึง 28 วัน ส่วนคลาส 1 เป็นกลุ่มของผู้ป่วยมารับการรักษาซ้ำตั้งแต่ 29 ถึง 90 วัน แบบจำลองนี้สามารถช่วยในการวางแผนการรักษาของแพทย์ ผลการศึกษาพบว่าข้อมูลมีค่าผิดปกติและมีความไม่สมดุลของคลาสในข้อมูลโดยมีจำนวนคลาสหนึ่งมากกว่าอีกคลาสหนึ่งเป็นจำนวนมาก ผู้วิจัยจึงได้ทำการแก้ปัญหาโดยคัดกรองเอาค่าผิดปกติออกด้วยเทคนิค Support Vector Machine และปรับความสมดุลของข้อมูลด้วยวิธี Synthetic Minority Over-sampling Technique (SMOTE) แล้วพัฒนาแบบจำลองด้วยเทคนิคต้นไม้ตัดสินใจ (C4.5) นาอีฟเบย์ (Naïve Bayes) และ PART decision list นอกจากนั้นผู้วิจัยยังได้ใช้ 10-fold cross validation ในการแบ่งข้อมูลออกเป็นชุดข้อมูลสอนและชุดข้อมูลทดสอบและได้ใช้ทำความถูกต้อง ค่าความไว และค่าความจำเพาะในการแสดงประสิทธิภาพในการพยากรณ์ของแบบจำลองอีกด้วย ผลการทดลองประสิทธิภาพในการพยากรณ์ของแบบจำลองพบว่า เทคนิค PART decision list สามารถพยากรณ์ได้ดีกว่าต้นไม้ตัดสินใจ (C4.5) และนาอีฟเบย์ (Naïve Bayes) ซึ่งมีค่าความ

ถูกต้อง (Accuracy) ร้อยละ 92.98 ค่าความไว (Sensitivity) ร้อยละ 92.95 และค่าความจำเพาะ (Specificity) ร้อยละ 93.02 ตามลำดับ

เชาวนันท์ โสโท และคณะ (2556) ทำการศึกษาเรื่องแบบจำลองการทำนายผลการรักษาผู้ป่วยมะเร็งปากมดลูกด้วยโครงข่ายประสาทเทียม โดยงานวิจัยนี้นำเสนอแบบจำลองเพื่อทำนายผลการรักษาผู้ป่วยมะเร็งปากมดลูกที่เข้ารับการรักษาด้วยวิธีการฉายรังสี โดยการประยุกต์ใช้โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ ในการศึกษาผู้วิจัยได้รวบรวมปัจจัยต่าง ๆ ที่มีความสัมพันธ์ต่อผลการรักษาจากงานวิจัยทางการแพทย์ที่เกี่ยวข้อง และได้ทำการวิเคราะห์ความสัมพันธ์ของปัจจัยด้วยวิธีแบบลำดับชั้น ผลการศึกษาพบว่าปัจจัยที่มีความสัมพันธ์ต่อผลการรักษาประกอบด้วย ระยะของโรค อายุ ขนาดของก้อนเนื้อ ประเภทของเซลล์มะเร็ง และน้ำหนักตัว จากการศึกษางานวิจัยที่เกี่ยวข้องพบว่าระดับฮีโมโกลบินมีผลตอบสนองต่อการรักษา ดังนั้นจึงได้เพิ่มระดับฮีโมโกลบินรวมกับปัจจัยที่ได้จากการวิเคราะห์รวมเป็น 6 ปัจจัย ในการสร้างแบบจำลองด้วยตัวแบบโครงข่ายประสาทเทียม ผลการศึกษาพบว่าตัวแบบจำลองที่นำเสนอถึงแม้จะมีประสิทธิภาพสูง แต่ค่าความจำเพาะมีค่าเป็น 0% ซึ่งจากการวิเคราะห์พบว่าปัญหาเกิดจากความไม่สมดุลของข้อมูลซึ่งมีข้อมูลกลุ่มหนึ่งมากกว่าอีกกลุ่มเป็นจำนวนมาก ดังนั้นผู้วิจัยจึงได้ทำการแก้ปัญหาปรับความไม่สมดุลของข้อมูลด้วยวิธี Cost-Sensitive Learning : CSL และวิธี Synthetic Minority Over-sampling Technique : SMOTE และทำการสร้างแบบจำลองข้อมูลโครงข่ายประสาทเทียมและแบบจำลองด้วยการถดถอยแบบโลจิสติก เมื่อเปรียบเทียบประสิทธิภาพการทำนายพบว่าโครงข่ายประสาทเทียมที่มีการแก้ปัญหาความไม่สมดุลของข้อมูลด้วยวิธีการสุ่มเพิ่มชุดข้อมูลตัวอย่างมีประสิทธิภาพการทำนายด้วยค่าความถูกต้อง 81.71% ค่าความไว 94.47% และค่าความจำเพาะ 55.47% สูงกว่าวิธีการถดถอยแบบโลจิสติกที่มีการแก้ปัญหาความไม่สมดุลของข้อมูลด้วยวิธี CSL ซึ่งมีค่าความถูกต้อง 81.00% ค่าความไว 84.52% และค่าความจำเพาะ 30.66%

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

Sa-ngasoongsong, A. and Chongwatpol, J. (2012) ได้ทำการวิจัยเรื่องปัจจัยเสี่ยงการเป็นโรคเบาหวานโดยใช้วิธีการทำเหมืองข้อมูล ในการศึกษานี้แบ่งกลุ่มคนไข้ออกเป็น 3 กลุ่ม โดยพิจารณาจากค่าใช้จ่าย ตรวจสอบตัวแบบในการวิเคราะห์ที่ซับซ้อนมากขึ้นโดยใช้เทคนิคการทำเหมืองข้อมูลโดยใช้โปรแกรม SAS Enterprise Miner™ 7.1 สามารถทำนายและอธิบายสาเหตุของการเป็นโรคเบาหวานเพิ่มขึ้นได้ดีในคนไข้ผู้ใหญ่ในแต่ละกลุ่มของค่าใช้จ่าย การวิเคราะห์เบื้องต้น

พบว่าความดันเลือดสูง อายุ คลอเรสเตอรอล ดัชนีมวลกาย รายได้ทั้งหมด เพศ การเป็นโรคหัวใจ สถานภาพสมรส การตรวจฟัน และการวินิจฉัยโรคหอบหืด เป็นปัจจัยเสี่ยงที่สำคัญ ถ้าพิจารณาตัวแบบทั้งหมดพบว่าวิธีการถดถอยโลจิสติกให้ผลอัตราการจำแนกกลุ่มผิดทั้งหมดดีที่สุด 22.89% แม้ว่าวิธีโครงข่ายประสาทเทียมมีอัตราความผิดพลาดเชิงลบต่ำที่สุดคือ 20.55% เรายังคงเลือกตัวแบบการถดถอยโลจิสติกเป็นตัวแบบสุดท้ายเพื่อทำนายคนไข้ที่เป็นโรคเบาหวาน หลังจากทำการถดถอยโลจิสติกด้วยวิธีขั้นตอน (stepwise) พบว่ามีปัจจัยเสี่ยง 15 ปัจจัยที่ใช้เพื่อทำนายความชุกของโรคเบาหวาน สำหรับกลุ่มค่าใช้จ่ายต่ำ (<5,750 ดอลลาร์) พบว่าวิธีแผนภาพต้นไม้เพื่อการตัดสินใจให้ผลอัตราการจำแนกกลุ่มผิดทั้งหมดดีที่สุด 24.56% รองลงมาคือวิธีการถดถอยโลจิสติกและโครงข่ายประสาทเทียมมีอัตราการจำแนกกลุ่มผิด 25.43% และ 25.68% ตามลำดับ การเลือกตัวแปรโดยวิธีขั้นตอนพบว่าตัวแปร 7 ตัว คือ อายุ คลอเรสเตอรอล ดัชนีมวลกาย ความดันเลือดสูง รายได้ทั้งหมด เพศ และการวินิจฉัยโรคหอบหืดเป็นปัจจัยสำคัญเพื่อทำนายการเป็นโรคเบาหวานในกลุ่มค่าใช้จ่ายนี้ สำหรับกลุ่มก่อนเป็นโรคเบาหวานหรือกลุ่มค่าใช้จ่ายปานกลาง (5,750-19,400 ดอลลาร์) พบว่าวิธีการถดถอยโลจิสติกให้ผลอัตราการจำแนกกลุ่มผิดทั้งหมดดีที่สุด 29.74% รองลงมาคือวิธีโครงข่ายประสาทเทียมและวิธีแผนภาพต้นไม้เพื่อการตัดสินใจตามลำดับ การเลือกตัวแปรโดยวิธีขั้นตอนพบว่าตัวแปร 8 ตัว คือ คลอเรสเตอรอล ดัชนีมวลกาย ความดันเลือดสูง องศาสูงสุด การเป็นไข้หวัด การเป็นโรคหัวใจ การเป็นโรคตาต้อกระจก และสถานภาพสมรสเป็นปัจจัยสำคัญเพื่อทำนายการเป็นโรคเบาหวานในกลุ่มค่าใช้จ่ายนี้ ส่วนกลุ่มเสี่ยงเป็นโรคเบาหวานหรือกลุ่มค่าใช้จ่ายสูง (> 19,400 ดอลลาร์) พบว่าวิธีแผนภาพต้นไม้เพื่อการตัดสินใจให้ผลอัตราการจำแนกกลุ่มผิดทั้งหมดดีที่สุด 41.80% รองลงมาคือวิธีโครงข่ายประสาทเทียมและวิธีการถดถอยโลจิสติกตามลำดับ การเลือกตัวแปรโดยวิธีขั้นตอนพบว่าตัวแปร 4 ตัว คือ ดัชนีมวลกาย อายุ การเป็นโรคตาต้อกระจก และสถานภาพสมรสเป็นปัจจัยสำคัญเพื่อทำนายการเป็นโรคเบาหวานในกลุ่มค่าใช้จ่ายนี้

Al-Rofiyee, A. et al. (2013) ศึกษาการใช้วิธีการทำนายในการทำเหมืองข้อมูลการวินิจฉัยโรคเบาหวาน โดยมุ่งเน้นการวิเคราะห์เชิงทำนายของการวินิจฉัยโรคเบาหวานโดยใช้วิธีโครงข่ายประสาทเทียม ใช้ซอฟแวร์ WEKA เป็นเครื่องมือสำหรับวินิจฉัยโรคเบาหวานและใช้ข้อมูลโรคเบาหวานจากฐานข้อมูล UCI สำหรับการวิเคราะห์ ชุดข้อมูลที่ทำการศึกษาและวิเคราะห์เพื่อหาตัวแบบที่มีประสิทธิภาพซึ่งทำนายและวินิจฉัยโรคเบาหวาน ได้แบ่งข้อมูลออกเป็น 3 ส่วน คือ ข้อมูลฝึกหัด 60% ข้อมูลทดสอบ 20% และข้อมูลประยุกต์ใช้ 20% ใช้อัตราการเรียนรู้ 0.3 และโมเมนตัม 0.2 พบว่าวิธีโครงข่ายประสาทเทียม อัลกอริทึมเพอร์เซปตรอนแบบหลายชั้น (Multilayer perceptron) ให้การทำนายชุดข้อมูลฝึกหัดมีความถูกต้องมากที่สุด 97.61% เมื่อพารามิเตอร์ชั้นซ่อนเป็น 20 และจำนวนรอบ 8,000 รอบ ส่วนชุดข้อมูลทดสอบและชุดข้อมูลประยุกต์ใช้ให้การทำนายมีความถูกต้องมากที่สุด 68.18% และ 81.8% เมื่อพารามิเตอร์ชั้นซ่อนเป็น 50 และจำนวนรอบ 500 รอบ

Cela, E. and Frasheri, N. (2013) ได้ทำการวิจัยเรื่องเทคนิคการทำเหมืองข้อมูลและเครื่องมือที่ใช้ในฐานข้อมูลศูนย์สุขภาพ วัตถุประสงค์ของการศึกษาคือเพื่อหาการประยุกต์ใช้ที่สำคัญและหาแนวโน้มของการวิจัยในศูนย์สุขภาพ ระเบียบวิธีที่ใช้ในงานวิจัยนี้คือการสำรวจของการตีพิมพ์และวารสารเมื่อเร็ว ๆ นี้ทางด้านวิทยาศาสตร์คอมพิวเตอร์ วิศวกรรม และศูนย์สุขภาพ

การวิเคราะห์พบว่างานวิจัยส่วนใหญ่ใช้อัลกอริทึมการทำเหมืองข้อมูล เช่น แผนภาพต้นไม้เพื่อการตัดสินใจ ซัพพอร์ตเวกเตอร์แมชชีน โครงข่ายประสาทเทียม โครงข่ายเบย์เซียนและนาอีฟ เบย์ การถดถอยโลจิสติก อัลกอริทึมพันธุศาสตร์ กฎ Fuzzy และกฎความสัมพันธ์ พบว่าระดับความถูกต้องสามารถยอมรับได้ จากชุดข้อมูลโรคหัวใจนำมาจากฐานข้อมูล UCI พบว่าวิธีโครงข่ายประสาทเทียม โดยใช้อัลกอริทึมเพอร์เซปตรอนหลายชั้นมีความถูกต้องมากที่สุด 100% รองลงมาคือวิธีแผนภาพต้นไม้เพื่อการตัดสินใจมีความถูกต้อง 99.66% อย่างไรก็ตาม ความถูกต้องอาจจะแปรผันตามจำนวนคุณลักษณะและจำนวนตัวอย่างในชุดข้อมูล ส่วนชุดข้อมูลโรคมะเร็งพบว่าวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ โดยใช้อัลกอริทึม C5.0 มีความถูกต้องมากที่สุด 100% รองลงมาคือวิธีนาอีฟ เบย์มีความถูกต้อง 95.0%

Folorunsho, O. (2013) ทำการเปรียบเทียบประสิทธิภาพเทคนิคการทำเหมืองข้อมูลที่แตกต่างกันในการค้นหาความรู้จากฐานข้อมูลทางการแพทย์ โดยฐานข้อมูลทางการแพทย์มีความสำคัญที่ใช้ในการทำนายสุขภาพของคนไข้ เพื่อให้ผลการทำนายที่ดีที่สุด เทคนิคจำเป็นต้องมีความถูกต้องสูง ในงานวิจัยนี้ เราสร้างตัวแบบจากคนไข้โรคเบาหวานและใช้เพื่อทำนายความน่าจะเป็นของการเกิดโรคเบาหวานของคนไข้ ทำการวัดประสิทธิภาพของอัลกอริทึมโครงข่ายประสาทเทียมและอัลกอริทึมแผนภาพต้นไม้เพื่อการตัดสินใจของข้อมูลทางการแพทย์ โดยพิจารณาจากระยะเวลาที่ใช้ในการสร้างตัวแบบ สถิติแคปป้า ความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) และความคลาดเคลื่อนกำลังสองสัมพัทธ์ (RSE) ทำการเลือกตัวจำแนกที่มีประสิทธิภาพสูงที่สุดและใช้สำหรับการทำนาย ผลการศึกษาพบว่าอัลกอริทึมแผนภาพต้นไม้เพื่อการตัดสินใจมีประสิทธิภาพสูงกว่าอัลกอริทึมโครงข่ายประสาทเทียม

Rahman, R. M. and Afroz, F. (2013) ทำการเปรียบเทียบเทคนิคการจำแนกกลุ่มโดยใช้เครื่องมือการทำเหมืองข้อมูลที่แตกต่างกันสำหรับการวินิจฉัยโรคเบาหวาน ซึ่งการไม่มีหลักฐานการวินิจฉัยทางการแพทย์ มีความยากสำหรับผู้เชี่ยวชาญเพื่อสรุปเอาเองเกี่ยวกับอาการของโรคด้วยการยืนยัน การทดสอบจำนวนมากโดยทั่วไปเกี่ยวข้องกับการจัดกลุ่มหรือการจำแนกกลุ่มของข้อมูลขนาดใหญ่ ในงานวิจัยนี้แสดงการเปรียบเทียบของเทคนิคการจำแนกกลุ่มต่าง ๆ โดยใช้เครื่องมือการทำเหมืองข้อมูล 3 วิธี คือ WEKA, TANAGRA และ MATLAB วัตถุประสงค์ของงานวิจัยนี้เพื่อวิเคราะห์ประสิทธิภาพของเทคนิคการจำแนกกลุ่มที่แตกต่างกันสำหรับชุดข้อมูลขนาดใหญ่ ข้อมูลโรคเบาหวานประกอบด้วยคนไข้ 768 ราย และ 9 คุณลักษณะ (8 สำหรับข้อมูลเข้าและ 1 สำหรับข้อมูลออก) จะใช้เพื่อทดสอบและแสดงให้เห็นถึงความแตกต่างระหว่างวิธีการจำแนกกลุ่ม ในการวัดและตรวจสอบประสิทธิภาพระหว่างวิธีจำแนกกลุ่มจะใช้วิธีโครงข่ายประสาทเทียม อัลกอริทึมเพอร์เซปตรอนแบบหลายชั้น (MLP) ตัวจำแนกโครงข่ายเบย์ J48graft (แผนภาพต้นไม้ C4.5 Revision 8) JRip (RIPPER) ตัวจำแนก Fuzzy Lattice Reasoning (FLR) ใช้โปรแกรม WEKA ส่วน MLP นาอีฟ เบย์ และแผนภาพต้นไม้เพื่อการตัดสินใจ C4.5 ใช้โปรแกรม TANAGRA และ Fuzzy Inference System (FIS) และ Adaptive Neuro-Fuzzy Inference System (ANFIS) ใช้โปรแกรม MATLAB โดยที่ข้อมูล 66% ใช้สำหรับฝึกหัด และข้อมูลที่เหลือ 34% ใช้สำหรับทดสอบวัตถุประสงค์ พบว่าวิธี J48graft มีค่าความถูกต้องมากที่สุด 81.33% ส่วนวิธี FLR มีค่าความถูกต้องต่ำที่สุด 51.43% จากการเปรียบเทียบความคลาดเคลื่อนในบรรดาตัวจำแนกที่แตกต่างกันในโปรแกรม WEKA พบว่าวิธี J48graft ดีที่สุด

รองลงมาคือวิธีเบย์เนท ส่วนวิธี MLP & JRip ปานกลาง ตัวจำแนก FLR ใช้ระยะเวลาในการสร้างตัวแบบน้อยที่สุดประมาณ 0.025 วินาที รองลงมาคือโครงข่ายเบย์ใช้ระยะเวลา 0.04 วินาที J48graft ใช้ระยะเวลา 0.135 วินาที ส่วนอัลกอริทึม MLP ใช้ระยะเวลานานที่สุด 63.13 วินาที จากการตรวจสอบความถูกต้องโดยใช้โปรแกรม TANAGRA สำหรับตัวจำแนก 3 ตัว คือ MLP นาอีฟ เบย์ และ C4.5 พบว่าตัวจำแนกนาอีฟ เบย์ดีที่สุดเมื่อเทียบกับตัวจำแนกอื่น ๆ เนื่องจากมีค่าความถูกต้อง 100% และใช้ระยะเวลาในการสร้างตัวแบบน้อยที่สุด 0.001 วินาที เมื่อเทียบกับตัวจำแนกอื่น ๆ และผลจากโปรแกรม MATLAB เนื่องจากจำนวนรอบและชุดข้อมูลเหมือนกัน ความคลาดเคลื่อนเฉลี่ยจึงคล้ายคลึงกัน แต่ค่าความถูกต้องเป็น 71.51% และ 78.79% สำหรับ FIS และ ANFIS ตามลำดับ สำหรับโครงข่ายที่แตกต่างกัน

Ananthapadmanaban, K. R. and Parthiban, G. (2014) ทำการศึกษาการทำนายการเปลี่ยนแปลงโรคเบาหวานโดยใช้เทคทิคการจำแนกกลุ่มในการทำเหมืองข้อมูล โดยโรคเบาหวานที่จอประสาทตาเป็นโรคเบาหวานที่ตาที่มีความสำคัญ มีสาเหตุมาจากโรคแทรกซ้อน เกิดขึ้นเมื่อหลอดเลือดในม่านตาเรตินำเสื่อมถอยลง ทำให้สูญเสียการมองเห็นถ้าไม่ได้ทำการตรวจสอบในระยะเริ่มแรก เทคนิคการทำเหมืองข้อมูลจำนวนมากมีวัตถุประสงค์ที่แตกต่างกันโดยขึ้นอยู่กับวัตถุประสงค์ของตัวแบบ ผลลัพธ์ของเทคนิคการจำแนกกลุ่มในการทำเหมืองข้อมูลเปรียบเทียบโดยใช้เครื่องมือ rapid miner โดยใช้วิธีนาอีฟ เบย์และวิธีซัพพอร์ตเวกเตอร์แมชชีนเพื่อทำนายการตรวจจับโรคเบาหวานที่ตาพบว่าวิธีนาอีฟ เบย์มีค่าความถูกต้องมากกว่าคือ 83.37% ส่วนวิธีซัพพอร์ตเวกเตอร์แมชชีนมีค่าความถูกต้องเพียง 64.91%

Giri, T. N. and Todamal, S. R. (2014) ทำวิจัยเรื่องวิธีการทำเหมืองข้อมูลสำหรับการวินิจฉัยโรคเบาหวาชนิดที่ 2 โดยในทางการแพทย์ แพทย์จำเป็นต้องมีข้อมูลสำหรับทำการตัดสินใจ ในทุกวันนี้วิธีการทำเหมืองข้อมูลใช้ในการวิจัยทางการแพทย์ของโรคเบาหวานและใช้สำหรับวินิจฉัยเพื่อให้ประสบความสำเร็จ ดังนั้นในงานวิจัยนี้เน้นการวิเคราะห์ข้อมูลโรคเบาหวานโดยใช้เทคนิคการทำเหมืองข้อมูล เช่น อัลกอริทึมนาอีฟ เบย์ J48(C4.5) JRip โครงข่ายประสาทเทียม แผนภาพต้นไม้เพื่อการตัดสินใจ ความใกล้เคียงกันมากที่สุด Fuzzy logic และอัลกอรทึมพันธุศาสตร์ พบว่าอัลกอริทึม J48 เป็นเทคนิคที่ดีที่สุดโดยมีค่าความถูกต้อง 100% รองลงมาคือนาอีฟ เบย์ มีค่าความถูกต้อง 98.85% นอกจากนี้ J48 ยังใช้ระยะเวลาน้อยที่สุดในการสร้างตัวแบบคือ 6.58 วินาที รองลงมาคือ JRip ใช้ระยะเวลา 7.65 วินาที

Kumar, S. and Gayathri, P. (2014) ศึกษาการวิเคราะห์โรคเบาหวานในผู้ใหญ่โดยใช้อัลกอริทึมการจำแนกกลุ่มการทำเหมืองข้อมูล โดยโรคเบาหวานเป็นหนึ่งของการท้าทายต่อสุขภาพที่รุนแรงที่สุดในประเทศที่กำลังพัฒนาและประเทศที่พัฒนาแล้ว ในทางการแพทย์การทำนายในการทำเหมืองข้อมูลใช้เพื่อวินิจฉัยโรคเบาหวานในระยะเริ่มแรกซึ่งช่วยแพทย์ในการรักษาได้ ในทุกวันนี้คนไข้ได้รับผลกระทบโดยโรคต่าง ๆ ซึ่งพัฒนาไปสู่โรคแทรกซ้อน โรคเบาหวานในผู้ใหญ่ซึ่งเป็นปัญหาต่อสุขภาพ ในงานวิจัยนี้ วิธีการทำเหมืองข้อมูลจำนวนมากใช้เพื่อช่วยแพทย์ในการตรวจสอบโรคในระยะเริ่มแรกซึ่งช่วยลดโอกาสในการเป็นโรคเบาหวานชนิดที่ 2 ได้ โดยใช้อัลกอริทึมแผนภาพต้นไม้เพื่อการตัดสินใจ 3 ชนิด คือ Simple Cart, J48 และ NB Tree ใช้กับข้อมูลโรคเบาหวานชนิดที่ 2 ใน WEKA และวัดประสิทธิภาพของอัลกอริทึมโดยพิจารณาจากระยะเวลาที่ใช้ในการสร้างต้นไม้และความถูกต้องในการจำแนกตัวอย่าง ผลการทดลองพบว่าอัลกอริทึม J48

มีประสิทธิภาพสูงสุดโดยมีความถูกต้อง 93.5% และระยะเวลาในการสร้างแผนภาพต้นไม้น้อยที่สุด 0.02 วินาที

Kumar, V. P. and Velide, L. (2014) ศึกษาวิธีการทำเหมืองข้อมูลสำหรับการทำนายและการรักษาโรคเบาหวาน โดยความก้าวหน้าในคอมพิวเตอร์ช่วยจัดการข้อมูลขนาดใหญ่ งานเพื่อการวิเคราะห์ข้อมูลเข้าและได้ข้อมูลที่จำเป็นซึ่งสามารถกระทำโดยใช้เทคนิคการทำเหมืองข้อมูลต่าง ๆ การวินิจฉัยโรคมีความสำคัญและเป็นงานที่น่าเบื่อในทางการแพทย์ ดังนั้นงานวิจัยที่นำเสนอเน้น การวิเคราะห์ข้อมูลโรคเบาหวานโดยใช้เทคนิคการทำเหมืองข้อมูลซึ่งประกอบด้วย วิธีนาอีฟ เบย์ J48 (C4.5) JRip โครงข่ายประสาทเทียม แผนภาพต้นไม้เพื่อการตัดสินใจ ความใกล้เคียงกันมากที่สุด (KNN) Fuzzy logic และอัลกอริทึมพันธุศาสตร์ โดยพิจารณาจากความถูกต้องและระยะเวลา คุณลักษณะที่เลือก 9 คุณลักษณะ ประกอบด้วยเพศ diastolic, B.P., ปริมาณกลูโคสในพลาสมา ความหนาของรอยพับผิวหนัง ดัชนีมวลกาย ชนิดโรคเบาหวาน ระยะเวลาตั้งครรภ์ ระดับอินซูลินในซีรัม 2 ชั่วโมง และความน่าจะเป็นที่จะเป็นโรคเบาหวาน พบว่าวิธี J48 (C4.5) ดีกว่าวิธีอื่น ๆ โดยมีค่าความถูกต้อง 100% วิธี J48 ง่ายและเป็นตัวจำแนกที่มีความถูกต้องเพื่อทำแผนภาพต้นไม้เพื่อการตัดสินใจดีกว่าตัวจำแนกอื่น ๆ วิธีความใกล้เคียงกันมากที่สุดใช้ระยะเวลานานในการวิเคราะห์ข้อมูล ส่วนวิธี J48 มีความถูกต้องสูงและใช้ระยะเวลาน้อยที่สุด สิ่งนี้แสดงว่าค่าใช้จ่ายในการคำนวณสำหรับการวิเคราะห์ข้อมูลต่ำสำหรับวิธี J48

Sanakal, R. and Jayakumari. S. T. (2014) ได้ทำวิจัยเรื่องการทำนายของโรคเบาหวานโดยใช้วิธีการทำเหมืองข้อมูลการจัดกลุ่ม Fuzzy C Mean (FCM) และวิธีซัพพอร์ตเวกเตอร์แมชชีน การศึกษานี้พยายามใช้วิธีการทำเหมืองข้อมูลเพื่อวิเคราะห์ข้อมูลโรคเบาหวานและการวินิจฉัยโรคเบาหวาน ข้อมูลทางการแพทย์นำมาจากฐานข้อมูล UCI ประกอบด้วยข้อมูลเข้า 9 ตัว และข้อมูลออก 1 ตัว ซึ่งระบุว่าคนไข้เป็นโรคเบาหวานหรือไม่ ข้อมูลทั้งหมดประกอบด้วย 768 คน การจัดกลุ่ม FCM โดยใช้โปรแกรม MATLAB R2012a อัลกอริทึมการจัดกลุ่มใช้ชุดข้อมูลฝึกหัด 576 คน และชุดข้อมูลทดสอบ 192 คน โดยพิจารณาจากค่าความถูกต้อง ค่า sensitivity และค่า specificity พบว่าผลที่ดีที่สุดคือวิธี FCM มีค่าความถูกต้อง 94.3% ค่า sensitivity 95.4% และค่า specificity 93.8% ส่วนวิธีซัพพอร์ตเวกเตอร์แมชชีนมีค่าความถูกต้องเพียง 59.5% ค่า sensitivity 77.4% และค่า specificity 26.1%

Agrawal, P. and Dewangan, A. K. (2015) ศึกษาการสำรวจโดยย่อของเทคนิคที่ใช้สำหรับการวินิจฉัยโรคเบาหวาน โดยโรคเบาหวานเป็นความยุ่งเหยิงของเมตาโบลิซึมซึ่งเกิดขึ้นจากการทำงานผิดพลาดในการหลั่งอินซูลิน โรคเบาหวานก่อให้เกิดความผิดปกติและความผิดพลาดของอวัยวะต่าง ๆ เช่น ไต ตา ประสาท หลอดเลือด และหัวใจ ในอดีตมีเทคนิคจำนวนมากในการตรวจสอบโรคเบาหวาน ส่วนในปัจจุบันการวินิจฉัยโรคเบาหวานมีความสำคัญมากโดยใช้เทคนิคต่าง ๆ เช่น วิธีความใกล้เคียงกันมากที่สุด โครงข่ายประสาทเทียม ซัพพอร์ตเวกเตอร์แมชชีน นาอีฟ เบย์ ID3, C4.5, C5.0 และ CART พบว่าวิธีซัพพอร์ตเวกเตร์แมชชีนมีความถูกต้องในการจำแนกกลุ่มสูงสุดคือ 81.77% รองลงมาคือวิธีโครงข่ายประสาทเทียม

Iyer, A. et al. (2015) ศึกษาการวินิจฉัยโรคเบาหวานโดยใช้เทคนิคการจำแนกกลุ่มในการทำเหมืองข้อมูล ในงานวิจัยนี้มีวัตถุประสงค์เพื่อหาแนวทางแก้ไขปัญหาเพื่อวินิจฉัยโรคโดย

วิเคราะห์รูปแบบที่พบในข้อมูลโดยใช้การวิเคราะห์จำแนกกลุ่มโดยใช้อัลกอริทึมแผนภาพต้นไม้เพื่อการตัดสินใจ (J48) และอัลกอริทึมนาอีฟ เบย์ งานวิจัยนี้หวังว่าจะเสนอเทคนิคที่มีประสิทธิภาพและมีความรวดเร็วในการวินิจฉัยโรคเพื่อช่วยรักษาคนไข้ได้ วิธี J48 ใช้ 2 เทคนิค คือเทคนิค cross-validation และ percentage split (70:30) โดยแบ่งชุดข้อมูลออกเป็น 70% สำหรับการฝึกหัด และ 30% สำหรับการทดสอบ ส่วนวิธีนาอีฟ เบย์ใช้เทคนิค percentage split (70:30) เพียงเทคนิคเดียว โดยพิจารณาจากความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Error : MAE) ความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Squared Error : MSE) ความคลาดเคลื่อนสัมบูรณ์สัมพัทธ์ (Relative Absolute Error : RAE) และความคลาดเคลื่อนกำลังสองสัมพัทธ์ (Relative Squared Error : RSE) พบว่าทั้ง 2 วิธี มีอัตราความคลาดเคลื่อนแตกต่างกันเล็กน้อยโดยใช้อัตราส่วน 70:30 สำหรับวิธีนาอีฟ เบย์ให้ความคลาดเคลื่อนน้อยกว่าวิธีแผนภาพต้นไม้ J48 ทั้ง 2 วิธี ตัวแบบทั้ง 2 ตัวแบบมีประสิทธิภาพในการวินิจฉัยโรคเบาหวานโดยใช้อัตราส่วน 70:30 ของชุดข้อมูล ตัวแบบที่พัฒนาสำหรับวินิจฉัยโรคเบาหวานจำเป็นต้องใช้ชุดข้อมูลฝึกหัดและชุดข้อมูลทดสอบ

Devi, M. R. (2016) ศึกษาการวิเคราะห์ของเทคนิคการทำเหมืองข้อมูลต่าง ๆ เพื่อทำนายโรคเบาหวาน โดยวิธีการทำเหมืองข้อมูลช่วยวินิจฉัยโรคของคนไข้ โรคเบาหวาน Mellitus เป็นโรคเรื้อรังที่ส่งผลต่ออวัยวะต่าง ๆ ของร่างกายมนุษย์ การทำนายเบื้องต้นช่วยยืดอายุของคนและสามารถควบคุมโรคได้ ในงานวิจัยนี้สำรวจการทำนายเบื้องต้นของโรคเบาหวานโดยใช้เทคนิคการทำเหมืองข้อมูลต่าง ๆ เช่น นาอีฟ เบย์, MLP, โครงข่ายเบย์เซียน, C4.5, Amalgam KNN, ANFIS, PLS-LDA, Homogeneity-Based, โครงข่ายประสาทเทียม (ANN) และแผนภาพต้นไม้เพื่อการตัดสินใจ J48 แบบดัดแปลง เป็นต้น ชุดข้อมูลประกอบด้วยคน 768 ราย จากชุดข้อมูล PIMA Indian Dataset เพื่อหาความถูกต้องของเทคนิคการทำเหมืองข้อมูลในการทำนาย การวิเคราะห์พบว่าวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ J48 แบบดัดแปลงมีค่าความถูกต้องมากกว่าวิธีการอื่น ๆ

บทที่ 3
วิธีการดำเนินงาน

3.1 อุปกรณ์ในการวิจัย

3.1.1 อุปกรณ์ที่มีอยู่แล้ว

1) เครื่องคอมพิวเตอร์
2) เครื่องพิมพ์เลเซอร์
3) โปรแกรมสำเร็จรูป Weka version 3.7
4) โปรแกรมสำเร็จรูป Microsoft Excel 2007

3.1.2 อุปกรณ์ที่ต้องการเพิ่ม

1) แบบฟอร์มใบเข้าตรวจรับการรักษา

3.2 ขั้นตอนการดำเนินงานวิจัย

งานวิจัยนี้ได้ทำการศึกษาเรื่องการเปรียบเทียบประสิทธิภาพในการทำนายผลการเป็นโรคเบาหวานของผู้มาตรวจร่างกายในโรงพยาบาล ระหว่างวิธีความใกล้เคียงกันมากที่สุด วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ วิธีโครงข่ายประสาท วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม และวิธีนาอีฟ เบย์ โดยมีขั้นตอนการดำเนินงานดังนี้

3.2.1 การเก็บรวบรวมข้อมูล

เก็บรวบรวมข้อมูลผลการวินิจฉัยของผู้มาตรวจร่างกายในโรงพยาบาล ตั้งแต่เดือนมกราคมถึงเดือนเมษายน พ.ศ. 2559 จำนวน 1,233 คน โดยข้อมูลประกอบด้วยคุณลักษณะ (Attribute) ต่าง ๆ ดังนี้

1. เพศ
    - 1.1 เพศชาย
    - 1.2 เพศหญิง
2. อายุ
3. น้ำหนัก
4. ส่วนสูง
5. ค่าดัชนีมวลกาย (BMI)
6. การออกกำลังกาย
    - 6.1 ไม่ได้ออกกำลังกาย
    - 6.2 ออกกำลังกาย
7. ความดันโลหิตสูง
    - 7.1 ไม่เป็น
    - 7.2 เป็น
8. ความเครียด
    - 8.1 ไม่มี
    - 8.2 มี

9. การดื่มเครื่องดื่มที่มีแอลกอฮอล์
    9.1 ไม่ได้ดื่ม
    9.2 ดื่ม
10. กรรมพันธุ์
    10.1 พ่อแม่ไม่เคยเป็นโรคเบาหวาน
    10.2 พ่อแม่เป็นโรคเบาหวาน
11. การรับประทานอาหาร
    11.1 หวาน
    11.2 มัน
    11.3 เค็ม
12. ผลการตรวจวินิจฉัยการเป็นโรคเบาหวาน
    12.1 ไม่เป็นโรคเบาหวาน
    12.2 เป็นโรคเบาหวาน

3.2.2 บันทึกข้อมูลผลการวินิจฉัยการเป็นโรคเบาหวานของผู้มาทำการตรวจร่างกายจำนวน 1,233 คน ลงในโปรแกรม Microsoft Excel โดยให้แนวคอลัมน์เป็นเพศ อายุ น้ำหนัก ส่วนสูง ค่าดัชนีมวลกาย การออกกำลังกาย ความดันโลหิตสูง ความเครียด การดื่มเครื่องดื่มที่มีแอลกอฮอล์ กรรมพันธุ์ การรับประทานอาหาร และผลการวินิจฉัยการเป็นโรคเบาหวานของผู้มาทำการตรวจร่างกาย ส่วนแนวแถวเป็นลำดับที่ของผู้มาตรวจร่างกายการเป็นโรคเบาหวาน

3.2.3 การแบ่งข้อมูล

นำข้อมูลทั้งหมดมาแบ่งออกเป็น 3 ส่วน จากข้อมูลการเป็นโรคเบาหวานของผู้ที่มาทำการตรวจร่างกาย จำนวน 1,233 คน เพื่อใช้สำหรับการวิเคราะห์ข้อมูล โดยแบ่งข้อมูลออกเป็นสัดส่วนดังนี้ (พยูน พาณิชย์กุล, 2548)

ส่วนที่ 1 : ข้อมูลชุดฝึกหัด (Training Data Set) เพื่อนำไปสร้างตัวแบบ
มีข้อมูล 70% ของข้อมูลทั้งหมด ซึ่งจะได้ข้อมูลในส่วนที่ 1 จำนวน 863 คน

ส่วนที่ 2 : ข้อมูลชุดทดสอบประสิทธิภาพ (Evaluation Data Set) เพื่อนำไปทดสอบความถูกต้องของตัวแบบ
มีข้อมูล 20% ของข้อมูลทั้งหมด ซึ่งจะได้ข้อมูลในส่วนที่ 2 จำนวน 247 คน

ส่วนที่ 3 : ข้อมูลชุดทดสอบ (Testing Data Set) เพื่อนำไปทำนายตัวแบบ
มีข้อมูล 10% ของข้อมูลทั้งหมด ซึ่งจะได้ข้อมูลในส่วนที่ 3 จำนวน 123 คน

3.2.4 เครื่องมือที่ใช้ในการวิจัย

ในการวิเคราะห์ข้อมูลครั้งนี้ใช้โปรแกรม WEKA (Waikato Environment for Knowledge Analysis) เวอร์ชั่น 3.7 ซึ่งเป็นโปรแกรมที่สามารถดาวน์โหลดได้จากเว็บไซต์ ซึ่งอยู่ภายใต้การควบคุมของ GPL License ซึ่งโปรแกรม WEKA ได้ถูกพัฒนามาจากภาษาจาวาทั้งหมด ซึ่งเป็นที่นิยมในการใช้งานด้านการทำเหมืองข้อมูล

3.2.5 วิธีการวิเคราะห์ข้อมูล

ในการทำวิจัยครั้งนี้ คณะผู้วิจัยได้จัดข้อมูลแต่ละชุดออกเป็น 3 ส่วน โดยส่วนที่ 1 ใช้ข้อมูล 70 เปอร์เซ็นต์ของข้อมูลทั้งหมดในการสร้างตัวแบบ ส่วนที่ 2 ใช้ข้อมูล 20 เปอร์เซ็นต์ของข้อมูลทั้งหมดในการทดสอบความถูกต้องของตัวแบบ ส่วนที่ 3 ใช้ข้อมูล 10 เปอร์เซ็นต์ของข้อมูลทั้งหมดในการทำนายตัวแบบ แปลงไฟล์ข้อมูลให้เป็นนามสกุล *.csv เพื่อใช้วิเคราะห์ประสิทธิภาพการจำแนกกลุ่มข้อมูลในโปรแกรม WEKA ซึ่งเป็นโปรแกรมที่สามารถนำมาทดสอบอัลกอริทึมของวิธีการจำแนกกลุ่มได้เนื่องจากมีอัลกอริทึมที่ได้ระบุไว้ให้เลือกใช้ในโปรแกรมครบตามที่กำหนด ผู้วิจัยได้กำหนดวิธีการจำแนกกลุ่มเพื่อนำมาทดสอบดังนี้

1) วิธีความใกล้เคียงกันมากที่สุด (K-Nearest Neighbor : KNN) ใช้อัลกอริทึมชนิด IBk เนื่องจากเป็นฟังก์ชันหลักที่สนใจ ซึ่งเป็นพื้นฐานของอัลกอริทึม 8.1 อัลกอริทึม IBk ยังสามารถกำหนดน้ำหนักระยะห่างและทางเลือก (option) เพื่อกำหนดค่า k โดยใช้ cross-validation (Kumar, V. and Wu, X., 2009)

2) วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ (Decision Tree) ใช้อัลกอริทึมชนิด J48 ซึ่งพัฒนามาจาก ID3 สามารถใช้ได้กับข้อมูลแบบไม่ต่อเนื่องและแบบต่อเนื่อง ต่างจาก ID3 ที่ใช้ได้เพียงข้อมูลแบบไม่ต่อเนื่องเท่านั้น (รุจิรา ธรรมสมบัติ, 2554)

3) วิธีโครงข่ายประสาท (Neural Network) ใช้อัลกอริทึมชนิดเพอร์เซปตรอนหลายชั้น (Multilayer Perceptron) โดยกำหนดค่าอัตราการเรียนรู้ (Learning Rate) เป็น 0.1, 0.2, 0.3, 0.4 และ 0.5 ค่าโมเมนตัม (Momentum) เป็น 0.5, 0.6, 0.7, 0.8 และ 0.9 จำนวนรอบการสอน (Training Time) 20,000 รอบ การวิจัยครั้งนี้ใช้อัลกอริทึมของวิธีโครงข่ายประสาทชนิดเพอร์เซปตรอนหลายชั้นที่มีชั้นซ่อน (Hidden Layer) 1 ชั้น แม้ว่าโครงสร้างโครงข่ายประสาทที่ซับซ้อนสามารถมีชั้นซ่อนมากกว่า 1 ชั้น แต่ในทางปฏิบัติการกำหนดชั้นซ่อน 1 ชั้น ก็เพียงพอต่อการวิเคราะห์ข้อมูล (Berson, A. and Stephen, J. Smith, 1997)

4) วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) ใช้อัลกอริทึม SMO ชนิดโพลิโนเมียลเคอร์เนล (Polynomial Kernel) เนื่องจากงานวิจัยที่อ้างอิงจากวาทินี นุ้ยเพียร และคณะ (2553) ได้ผลว่าวิธีซัพพอร์ตเวกเตอร์แมชชีนที่ใช้อัลกอริทึมชนิดโพลิโนเมียลเคอร์เนลดีที่สุด

5) วิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม (Binary Logistic Regression Method) เป็นการวิเคราะห์การถดถอยแบบหนึ่งโดยที่ตัวแปรตามเป็นตัวแปรเชิงคุณภาพ ส่วนตัวแปรอิสระอาจจะเป็นตัวแปรเชิงปริมาณหรือเชิงคุณภาพ หรืออาจจะมีทั้งตัวแปรเชิงปริมาณและตัวแปรเชิงคุณภาพก็ได้ (กัลยา วานิชย์บัญชา, 2552)

6) วิธีนาอีฟ เบย์ (Naïve Bayes Method) คืออัลกอริทึมที่ใช้หลักการของความน่าจะเป็นในการคัดกรองแต่ละคำตอบ (Class) โดยมีคำตอบ 2 คำตอบ (สายชล สินสมบูรณ์ทอง, 2558)

การนำผลการวิเคราะห์ข้อมูลมาประเมินผลเพื่อเปรียบเทียบประสิทธิภาพในการทำนายผลการเป็นโรคเบาหวานของผู้มาทำตรวจร่างกาย ใช้วิธีการจำแนกกลุ่ม 6 วิธี โดยนำข้อมูลที่ได้จากการวิเคราะห์มาเปรียบเทียบว่าวิธีในการจำแนกกลุ่มข้อมูลวิธีใดระหว่างวิธีความใกล้เคียงกันมากที่สุด วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ วิธีโครงข่ายประสาท วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีการถดถอยโลจิสติก และวิธีนาอีฟ เบย์ ที่มีค่าความถูกต้อง (Accuracy) ค่าอัตราความถูกต้องเชิงบวก (TP Rate) ค่าอัตราความถูกต้องเชิงลบ (TN Rate) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าความถ่วงดุล (F-Measure) มากที่สุด ส่วนค่าอัตราความผิดพลาดเชิงบวก (FP Rate) ค่าอัตราความผิดพลาดเชิงลบ (FN Rate) ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Error : MSE) และค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error : MSE) ที่มีค่าน้อยที่สุด จะทำให้มีประสิทธิภาพในการทำนายผลดีที่สุด

บทที่ 4

ผลการวิเคราะห์ข้อมูล

4.1 การวิเคราะห์ข้อมูล

จากโปรแกรม Weka 3.7 แสดงผลลัพธ์ของการทำนายการเป็นโรคเบาหวานด้วยวิธีความใกล้เคียงกันมากที่สุด วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ วิธีโครงข่ายประสาท วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม และวิธีนาอีฟ เบย์ ได้ผลดังนี้

4.1.1 วิธีความใกล้เคียงกันมากที่สุด (K-Nearest Neighbor Method)

วิธีความใกล้เคียงกันมากที่สุดนั้นมีอัลกอริทึมหลายอัลกอริทึม แต่ในงานวิจัยนี้ใช้อัลกอริทึม IBk ในการทำนายการเป็นโรคเบาหวาน ได้ผลดังนี้

4.1.1.1 การสร้างตัวแบบการเป็นโรคเบาหวาน

สร้างตัวแบบการเป็นโรคเบาหวานจากข้อมูลฝึกหัด (Training data set) ด้วยไฟล์ Diabetes-70.csv จำนวน 863 คน ได้ผลดังนี้

ตารางที่ 4.1 ผลในส่วนของการสรุปผล (Summary) จากการสร้างตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีความใกล้เคียงกันมากที่สุด

Correctly Classified Instances	746	86.4426 %
Incorrectly Classified Instances	117	13.5574 %
Kappa statistic	0.0729	
Mean absolute error	0.1365	
Root mean squared error	0.3677	
Relative absolute error	86.5737 %	
Root relative squared error	131.3309 %	
Coverage of cases (0.95 level)	86.4426 %	
Mean rel. region size (0.95 level)	50 %	
Total Number of Instances	863	

จากตารางที่ 4.1 จากข้อมูลจำนวน 863 คน ทำนายข้อมูลถูกต้องจำนวน 746 คน คิดเป็น 86.4426% และทำนายข้อมูลไม่ถูกต้องจำนวน 117 คน คิดเป็น 13.5574% มีค่าสถิติแคปปา (Kappa statistic) คือ 0.0729 แสดงว่าข้อมูลมีความสอดคล้องกันน้อยมาก มีค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) คือ 0.1365 ซึ่งมีค่าค่อนข้างน้อย แสดงว่าค่าที่ทำนายได้ใกล้เคียงกับค่าที่แท้จริงพอสมควร และมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) คือ $(0.3677)^2$ = 0.1352 ซึ่งมีค่าค่อนข้างน้อย แสดงว่าตัวแบบมีความถูกต้องพอสมควร

ตารางที่ 4.2 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบ (Detailed Accuracy By Class) จากการสร้างตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีความใกล้เคียงกันมากที่สุด

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.933	0.865	0.92	0.933	0.926	Negative
	TN Rate	FN Rate	Precision	Recall	F-Measure	Class
	0.135	0.067	0.159	0.135	0.146	Positive
Weighted Avg.	0.864	0.796	0.855	0.864	0.859	

จากตารางที่ 4.2 สำหรับคำตอบไม่เป็นโรคเบาหวาน (Class = Negative) มีค่าอัตราความถูกต้องเชิงบวก (TP Rate) = 0.933 ค่าอัตราความผิดพลาดเชิงบวก (FP Rate) = 0.865 ค่าความแม่นยำ (Precision) = 0.92 ค่าความระลึก (Recall) = 0.933 และค่าความถ่วงดุล (F-Measure) = 0.926 ส่วนคำตอบเป็นโรคเบาหวาน (Class = Positive) มีค่าอัตราความถูกต้องเชิงลบ (TN Rate) = 0.135 ค่าอัตราความผิดพลาดเชิงลบ (FN Rate) = 0.067 ค่าความแม่นยำ = 0.159 ค่าความระลึก = 0.135 และค่าความถ่วงดุล = 0.146

ตารางที่ 4.3 ผลในส่วนของเมตริกซ์ความสับสน (Confusion Matrix) จากการสร้างตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีความใกล้เคียงกันมากที่สุด

		ผลการจำแนกเป็นโรคเบาหวาน	
		ไม่เป็น	เป็น
ค่าที่แท้จริง	ไม่เป็น	736	53
	เป็น	64	10

จากตารางที่ 4.3 มีข้อมูล 863 คน ตัวแบบสามารถทำนายข้อมูลได้ถูกต้อง 746 คน โดยมีจำนวนข้อมูลที่จำแนกถูกว่าไม่เป็นโรคเบาหวาน 736 คน และจำนวนข้อมูลที่จำแนกถูกว่าเป็นโรคเบาหวาน 10 คน ตัวแบบทำนายข้อมูลไม่ถูกต้อง 117 คน โดยมีจำนวนข้อมูลที่จำแนกผิดว่าเป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วไม่เป็นโรคเบาหวาน 53 คน และจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วเป็นโรคเบาหวาน 64 คน

4.1.1.2 การทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน

ทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวานจากชุดข้อมูลทดสอบประสิทธิภาพ (Evaluation data set) ด้วยไฟล์ Diabetes-20.csv จำนวน 247 คน ได้ผลดังนี้

ตารางที่ 4.4 ผลในส่วนของการสรุปผลจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีความใกล้เคียงกันมากที่สุด

Correctly Classified Instances	219	88.664 %
Incorrectly Classified Instances	28	11.336 %
Kappa statistic	0.0646	
Mean absolute error	0.1142	
Root mean squared error	0.3363	
Coverage of cases (0.95 level)	88.664 %	
Total Number of Instances	247	

จากตารางที่ 4.4 จากข้อมูลจำนวน 247 คน ทำนายข้อมูลถูกต้องจำนวน 219 คน คิดเป็น 88.664% และทำนายข้อมูลไม่ถูกต้องจำนวน 28 คน คิดเป็น 11.336% มีค่าสถิติแคปปา (Kappa statistic) คือ 0.0646 แสดงว่าข้อมูลมีความสอดคล้องกันน้อยมาก มีค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) คือ 0.1142 ซึ่งมีค่าค่อนข้างน้อย แสดงว่าค่าที่ทำนายได้ใกล้เคียงกับค่าที่แท้จริงพอสมควร และมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) คือ $(0.3363)^2 = 0.1131$ ซึ่งมีค่าค่อนข้างน้อย แสดงว่าตัวแบบมีความถูกต้องพอสมควร

ตารางที่ 4.5 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีความใกล้เคียงกันมากที่สุด

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.943	0.882	0.935	0.943	0.939	Negative
	TN Rate	FN Rate	Precision	Recall	F-Measure	Class
	0.118	0.057	0.133	0.118	0.125	Positive
Weighted Avg.	0.887	0.826	0.88	0.887	0.883	

จากตารางที่ 4.5 สำหรับคำตอบไม่เป็นโรคเบาหวาน (Class = Negative) มีค่าอัตราความถูกต้องเชิงบวก = 0.943 ค่าอัตราความผิดพลาดเชิงบวก = 0.882 ค่าความแม่นยำ = 0.935 ค่าความระลึก = 0.943 และค่าความถ่วงดุล = 0.939 ส่วนคำตอบเป็นโรคเบาหวาน (Class = Positive) มีค่าอัตราความถูกต้องเชิงลบ = 0.118 ค่าอัตราความผิดพลาดเชิงลบ = 0.057 ค่าความแม่นยำ = 0.133 ค่าความระลึก = 0.118 และค่าความถ่วงดุล = 0.125

ตารางที่ 4.6 ผลในส่วนของเมตริกซ์ความสับสนจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีความใกล้เคียงกันมากที่สุด

		ผลการจำแนกเป็นโรคเบาหวาน	
		ไม่เป็น	เป็น
ค่าที่แท้จริง	ไม่เป็น	217	13
	เป็น	15	2

จากตารางที่ 4.6 มีข้อมูล 247 คน ตัวแบบสามารถทำนายข้อมูลได้ถูกต้อง 219 คน โดยมีจำนวนข้อมูลที่จำแนกถูกว่าไม่เป็นโรคเบาหวาน 217 คน และจำนวนข้อมูลที่จำแนกถูกว่าเป็นโรคเบาหวาน 2 คน ตัวแบบทำนายข้อมูลไม่ถูกต้อง 28 คน โดยมีจำนวนข้อมูลที่จำแนกผิดว่าเป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วไม่เป็นโรคเบาหวาน 13 คน และจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วเป็นโรคเบาหวาน 15 คน

4.1.1.3 การทำนายการเป็นโรคเบาหวาน

การทำนายการเป็นโรคเบาหวานจากชุดข้อมูลทดสอบ (Testing data set) ด้วยไฟล์ Diabetes-10.csv จำนวน 123 คน ได้ผลดังนี้

ตารางที่ 4.7 ผลในส่วนของการทำนายชุดข้อมูลทดสอบ (Predictions on test set) ด้วยวิธีความใกล้เคียงกันมากที่สุด

ตัวอย่าง (instance)	ค่าที่แท้จริง (actual)	ค่าที่ทำนายได้ (predicted)	ความคลาดเคลื่อน (error)	การทำนาย (prediction)
1	1:Negative	2:Positive	+	0.999
2	1:Negative	1:Negative		0.999
3	2:Positive	1:Negative	+	0.999
4	1:Negative	1:Negative		0.999
5	1:Negative	1:Negative		0.999
6	2:Positive	1:Negative	+	0.999
7	1:Negative	1:Negative		0.999
8	1:Negative	1:Negative		0.999
9	1:Negative	1:Negative		0.999
10	1:Negative	1:Negative		0.999
11	1:Negative	2:Positive	+	0.999
12	1:Negative	1:Negative		0.999
13	1:Negative	1:Negative		0.999
14	1:Negative	1:Negative		0.999
15	1:Negative	1:Negative		0.999
16	1:Negative	1:Negative		0.999
17	1:Negative	1:Negative		0.999
18	1:Negative	1:Negative		0.999
19	1:Negative	1:Negative		0.999
20	1:Negative	1:Negative		0.999
21	1:Negative	1:Negative		0.999
22	1:Negative	1:Negative		0.999
23	1:Negative	1:Negative		0.999
24	1:Negative	1:Negative		0.999
25	1:Negative	1:Negative		0.999
26	1:Negative	1:Negative		0.999
27	1:Negative	1:Negative		0.999
28	1:Negative	1:Negative		0.999
29	1:Negative	1:Negative		0.999
30	1:Negative	1:Negative		0.999
31	1:Negative	1:Negative		0.999
32	1:Negative	1:Negative		0.999
33	1:Negative	2:Positive	+	0.999
34	1:Negative	2:Positive	+	0.999
35	1:Negative	1:Negative		0.999
36	1:Negative	1:Negative		0.999
37	1:Negative	1:Negative		0.999
38	1:Negative	1:Negative		0.999
39	1:Negative	1:Negative		0.999
40	1:Negative	1:Negative		0.999
41	1:Negative	1:Negative		0.999
42	1:Negative	1:Negative		0.999
43	1:Negative	1:Negative		0.999
44	1:Negative	1:Negative		0.999
45	1:Negative	1:Negative		0.999
46	1:Negative	1:Negative		0.999

ตารางที่ 4.7 ผลในส่วนของการทำนายชุดข้อมูลทดสอบ (Predictions on test set) ด้วยวิธีความใกล้เคียงกันมากที่สุด (ต่อ)

ตัวอย่าง (instance)	ค่าที่แท้จริง (actual)	ค่าที่ทำนายได้ (predicted)	ความคลาดเคลื่อน (error)	การทำนาย (prediction)
47	2:Positive	1:Negative	+	0.999
48	1:Negative	1:Negative		0.999
49	1:Negative	1:Negative		0.999
50	1:Negative	1:Negative		0.999
51	1:Negative	1:Negative		0.999
52	1:Negative	1:Negative		0.999
53	1:Negative	1:Negative		0.999
54	1:Negative	1:Negative		0.999
55	1:Negative	1:Negative		0.999
56	1:Negative	1:Negative		0.999
57	1:Negative	1:Negative		0.999
58	1:Negative	1:Negative		0.999
59	1:Negative	1:Negative		0.999
60	1:Negative	1:Negative		0.999
61	2:Positive	1:Negative	+	0.999
62	1:Negative	1:Negative		0.999
63	1:Negative	1:Negative		0.999
64	1:Negative	1:Negative		0.999
65	2:Positive	1:Negative	+	0.999
66	1:Negative	1:Negative		0.999
67	1:Negative	2:Positive	+	0.999
68	1:Negative	1:Negative		0.999
69	2:Positive	1:Negative	+	0.999
70	1:Negative	1:Negative		0.999
71	1:Negative	1:Negative		0.999
72	1:Negative	1:Negative		0.999
73	1:Negative	1:Negative		0.999

ตัวอย่าง (instance)	ค่าที่แท้จริง (actual)	ค่าที่ทำนายได้ (predicted)	ความคลาดเคลื่อน (error)	การทำนาย (prediction)
74	2:Positive	1:Negative	+	0.999
75	1:Negative	1:Negative		0.999
76	1:Negative	1:Negative		0.999
77	1:Negative	1:Negative		0.999
78	1:Negative	1:Negative		0.999
79	1:Negative	1:Negative		0.999
80	1:Negative	1:Negative		0.999
81	1:Negative	1:Negative		0.999
82	2:Positive	1:Negative	+	0.999
83	1:Negative	1:Negative		0.999
84	1:Negative	1:Negative		0.999
85	1:Negative	1:Negative		0.999
86	1:Negative	1:Negative		0.999
87	1:Negative	1:Negative		0.999
88	1:Negative	1:Negative		0.999
89	1:Negative	1:Negative		0.999
90	1:Negative	1:Negative		0.999
91	1:Negative	1:Negative		0.999
92	1:Negative	1:Negative		0.999
93	1:Negative	1:Negative		0.999
94	1:Negative	1:Negative		0.999
95	1:Negative	1:Negative		0.999
96	1:Negative	2:Positive	+	0.999
97	1:Negative	1:Negative		0.999
98	1:Negative	1:Negative		0.999
99	1:Negative	1:Negative		0.999
100	1:Negative	1:Negative		0.999

ตารางที่ 4.7 ผลในส่วนของการทำนายชุดข้อมูลทดสอบ (Predictions on test set) ด้วยวิธีความใกล้เคียงกันมากที่สุด (ต่อ)

ตัวอย่าง nstance)	ค่าที่แท้จริง (actual)	ค่าที่ทำนายได้ (predicted)	ความคลาดเคลื่อน (error)	การทำนาย (prediction)
101	1:Negative	1:Negative		0.999
102	1:Negative	1:Negative		0.999
103	1:Negative	1:Negative		0.999
104	1:Negative	1:Negative		0.999
105	1:Negative	1:Negative		0.999
106	1:Negative	1:Negative		0.999
107	1:Negative	1:Negative		0.999
108	1:Negative	1:Negative		0.999
109	1:Negative	1:Negative		0.999
110	1:Negative	1:Negative		0.999
111	1:Negative	1:Negative		0.999
112	1:Negative	1:Negative		0.999

ตัวอย่าง (instance)	ค่าที่แท้จริง (actual)	ค่าที่ทำนายได้ (predicted)	ความคลาดเคลื่อน (error)	การทำนาย (prediction)
113	1:Negative	1:Negative		0.999
114	1:Negative	1:Negative		0.999
115	1:Negative	1:Negative		0.999
116	1:Negative	1:Negative		0.999
117	1:Negative	1:Negative		0.999
118	1:Negative	1:Negative		0.999
119	1:Negative	1:Negative		0.999
120	1:Negative	1:Negative		0.999
121	1:Negative	1:Negative		0.999
122	1:Negative	1:Negative		0.999
123	1:Negative	2:Positive	+	0.999

จากตารางที่ 4.7 ค่าทำนายที่ทายผิดในช่อง error จะมีเครื่องหมาย + อยู่จำนวน 15 ค่า คือค่าที่ 1, 3, 6, 11, 33, 34, 47, 61, 65, 67, 69, 74, 82, 96, 123 โดยที่ค่าที่ 3, 6, 47, 61, 65, 69, 74 และ 82 เป็นโรคเบาหวาน (class 2 : Positive) แต่ทำนายว่าไม่เป็นโรคเบาหวาน (class 1 : Negative) และค่าที่ 1, 11, 33, 34, 67, 96 และ 123 ไม่เป็นโรคเบาหวาน แต่ทำนายว่าเป็นโรคเบาหวาน

ตารางที่ 4.8 ผลในส่วนของการสรุปผลจากการทำนายการเป็นโรคเบาหวาน ด้วยวิธีความใกล้เคียงกันมากที่สุด

Correctly Classified Instances	108	87.8049 %
Incorrectly Classified Instances	15	12.1951 %
Kappa statistic	-0.0646	
Mean absolute error	0.1228	
Root mean squared error	0.3488	
Coverage of cases (0.95 level)	87.8049 %	
Total Number of Instances	123	

จากตารางที่ 4.8 จากข้อมูลจำนวน 123 คน ทำนายข้อมูลถูกต้องจำนวน 108 คน คิดเป็น 87.8049% และทำนายข้อมูลไม่ถูกต้องจำนวน 15 คน คิดเป็น 12.1951% มีค่าสถิติแคปปา (Kappa statistic) คือ -0.0646 แสดงว่าข้อมูลมีความสอดคล้องกันน้อยมาก มีค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) คือ 0.1228 ซึ่งมีค่าค่อนข้างน้อย แสดงว่าค่าที่ทำนายได้ใกล้เคียงกับค่าที่แท้จริงพอสมควร และมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) คือ $(0.3488)^2$ = 0.1217 ซึ่งมีค่าค่อนข้างน้อย แสดงว่าตัวแบบมีความถูกต้องพอสมควร

ตารางที่ 4.9 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการทำนายการเป็นโรคเบาหวาน ด้วยวิธีความใกล้เคียงกันมากที่สุด

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.939	1	0.931	0.939	0.935	Negative
	TN Rate	FN Rate	Precision	Recall	F-Measure	Class
	0	0.061	0	0	0	Positive
Weighted Avg.	0.878	0.939	0.87	0.878	0.874	

จากตารางที่ 4.9 สำหรับคำตอบไม่เป็นโรคเบาหวาน (Class = Negative) มีค่าอัตราความถูกต้องเชิงบวก = 0.939 ค่าอัตราความผิดพลาดเชิงบวก = 1 ค่าความแม่นยำ = 0.931 ค่าความระลึก = 0.939 และค่าความถ่วงดุล = 0.935 ส่วนคำตอบเป็นโรคเบาหวาน (Class = Positive) มีค่าอัตราความถูกต้องเชิงลบ = 0 ค่าอัตราความผิดพลาดเชิงลบ = 0.061 ค่าความแม่นยำ = 0 ค่าความระลึก = 0 และค่าความถ่วงดุล = 0

ตารางที่ 4.10 ผลในส่วนของเมตริกซ์ความสับสนจากการทำนายการเป็นโรคเบาหวาน ด้วยวิธีความใกล้เคียงกันมากที่สุด

		ผลการจำแนกเป็นโรคเบาหวาน	
		ไม่เป็น	เป็น
ค่าที่แท้จริง	ไม่เป็น	108	7
	เป็น	8	0

จากตารางที่ 4.10 มีข้อมูล 123 คน ตัวแบบสามารถทำนายข้อมูลได้ถูกต้อง 108 คน โดยมีจำนวนข้อมูลที่จำแนกถูกว่าไม่เป็นโรคเบาหวาน 108 คน และจำนวนข้อมูลที่จำแนกถูกว่าเป็นโรคเบาหวาน 0 คน ตัวแบบทำนายข้อมูลไม่ถูกต้อง 15 คน โดยมีจำนวนข้อมูลที่จำแนกผิดว่าเป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วเป็นโรคเบาหวาน 7 คน และจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วเป็นโรคเบาหวาน 8 คน

4.1.2 วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ (Decision Tree Method)

วิธีแผนภาพต้นไม้เพื่อการตัดสินใจนั้นมีอัลกอริทึมหลายอัลกอริทึม แต่ในงานวิจัยนี้ใช้อัลกอริทึม J48 ในการทำนายการเป็นโรคเบาหวาน ได้ผลดังนี้

4.1.2.1 การสร้างตัวแบบการเป็นโรคเบาหวาน

สร้างตัวแบบการเป็นโรคเบาหวานจากชุดข้อมูลฝึกหัด (Training data set) ด้วยไฟล์ Diabetes-70.csv จำนวน 863 คน ได้ผลดังนี้

ตารางที่ 4.11 ผลในส่วนของการสรุปผลจากการสร้างตัวแบบการเป็นโรคเบาหวาน

ด้วยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ

Correctly Classified Instances	784	90.8459 %
Incorrectly Classified Instances	79	9.1541 %
Kappa statistic	0.01	
Mean absolute error	0.1576	
Root mean squared error	0.2902	
Relative absolute error	99.9541 %	
Root relative squared error	103.6331%	
Coverage of cases (0.95 level)	98.6095 %	
Mean rel. region size (0.95 level)	91.8308 %	
Total Number of Instances	863	

จากตารางที่ 4.11 จากข้อมูลจำนวน 863 คน ทำนายข้อมูลถูกต้องจำนวน 784 คน คิดเป็น 90.8459% และทำนายข้อมูลไม่ถูกต้องจำนวน 79 คน คิดเป็น 9.1541% มีค่าสถิติแคปปา (Kappa statistic) คือ 0.01 แสดงว่าข้อมูลมีความสอดคล้องกันน้อยมาก มีค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) คือ 0.1576 ซึ่งมีค่าค่อนข้างน้อย แสดงว่าค่าที่ทำนายได้ใกล้เคียงกับค่าที่แท้จริงพอสมควร และมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) คือ $(0.2902)^2 = 0.0842$ ซึ่งมีค่าน้อยมาก แสดงว่าตัวแบบมีความถูกต้องดี

ตารางที่ 4.12 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการสร้างตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.992	0.986	0.915	0.992	0.952	Negative
	TN Rate	FN Rate	Precision	Recall	F-Measure	Class
	0.014	0.008	0.143	0.014	0.025	Positive
Weighted Avg.	0.908	0.903	0.849	0.908	0.872	

จากตารางที่ 4.12 สำหรับคำตอบไม่เป็นโรคเบาหวาน (Class = Negative) มีค่าอัตราความถูกต้องเชิงบวก = 0.992 ค่าอัตราความผิดพลาดเชิงบวก = 0.986 ค่าความแม่นยำ = 0.915 ค่าความระลึก = 0.992 และค่าความถ่วงดุล = 0.952 ส่วนคำตอบเป็นโรคเบาหวาน (Class = Positive) มีค่าอัตราความถูกต้องเชิงลบ = 0.014 ค่าอัตราความผิดพลาดเชิงลบ = 0.008 ค่าความแม่นยำ = 0.143 ค่าความระลึก = 0.014 และค่าความถ่วงดุล = 0.025

ตารางที่ 4.13 ผลในส่วนของเมตริกซ์ความสับสนจากการสร้างตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ

		ผลการจำแนกเป็นโรคเบาหวาน	
		ไม่เป็น	เป็น
ค่าที่แท้จริง	ไม่เป็น	783	6
	เป็น	73	1

จากตารางที่ 4.13 มีข้อมูล 863 คน ตัวแบบสามารถทำนายข้อมูลได้ถูกต้อง 784 คน โดยมีจำนวนข้อมูลที่จำแนกถูกว่าไม่เป็นโรคเบาหวาน 783 คน และจำนวนข้อมูลที่จำแนกถูกว่าเป็นโรคเบาหวาน 1 คน ตัวแบบทำนายข้อมูลไม่ถูกต้อง 79 คน โดยมีจำนวนข้อมูลที่จำแนกผิดว่าเป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วไม่เป็นโรคเบาหวาน 6 คน และจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วเป็นโรคเบาหวาน 73 คน

4.1.2.2 การทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน

ทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวานจากชุดข้อมูลทดสอบประสิทธิภาพ ด้วยไฟล์ Diabetes-20.csv จำนวน 247 คน ได้ผลดังนี้

ตารางที่ 4.14 ผลในส่วนของการสรุปผลจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ

Correctly Classified Instances	229	92.7126 %
Incorrectly Classified Instances	18	7.2874 %
Kappa statistic	-0.0077	
Mean absolute error	0.1328	
Root mean squared error	0.2633	
Coverage of cases (0.95 level)	95.9514 %	
Total Number of Instances	247	

จากตารางที่ 4.14 จากข้อมูลจำนวน 247 คน ทำนายข้อมูลถูกต้องจำนวน 229 คน คิดเป็น 92.7126% และทำนายข้อมูลไม่ถูกต้องจำนวน 18 คน คิดเป็น 7.2874% มีค่าสถิติแคปปา (Kappa statistic) คือ -0.0077 แสดงว่าข้อมูลมีความสอดคล้องกันน้อยมาก มีค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) คือ 0.1328 ซึ่งมีค่าค่อนข้างน้อย แสดงว่าค่าที่ทำนายได้ใกล้เคียงกับค่าที่แท้จริงพอสมควร และมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) คือ $(0.2633)^2 = 0.0693$ ซึ่งมีค่าน้อยมาก แสดงว่าตัวแบบมีความถูกต้องดี

ตารางที่ 4.15 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.996	1	0.931	0.996	0.962	Negative
	TN Rate	FN Rate	Precision	Recall	F-Measure	Class
	0	0.004	0	0	0	Positive
Weighted Avg.	0.927	0.931	0.867	0.927	0.896	

จากตารางที่ 4.15 สำหรับคำตอบไม่เป็นโรคเบาหวาน (Class = Negative) มีค่าอัตราความถูกต้องเชิงบวก = 0.996 ค่าอัตราความผิดพลาดเชิงบวก = 1 ค่าความแม่นยำ = 0.931 ค่าความระลึก = 0.996 และค่าความถ่วงดุล = 0.962 ส่วนคำตอบเป็นโรคเบาหวาน (Class = Positive) มีค่าอัตราความถูกต้องเชิงลบ = 0 ค่าอัตราความผิดพลาดเชิงลบ = 0.004 ค่าความแม่นยำ = 0 ค่าความระลึก = 0 และค่าความถ่วงดุล = 0

ตารางที่ 4.16 ผลในส่วนของเมตริกซ์ความสับสนจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ

		ผลการจำแนกเป็นโรคเบาหวาน	
		ไม่เป็น	เป็น
ค่าที่แท้จริง	ไม่เป็น	229	1
	เป็น	17	0

จากตารางที่ 4.16 มีข้อมูล 247 คน ตัวแบบสามารถทำนายข้อมูลได้ถูกต้อง 229 คน โดยมีจำนวนข้อมูลที่จำแนกถูกว่าไม่เป็นโรคเบาหวาน 229 คน และจำนวนข้อมูลที่จำแนกถูกว่าเป็นโรคเบาหวาน 0 คน ตัวแบบทำนายข้อมูลไม่ถูกต้อง 18 คน โดยมีจำนวนข้อมูลที่จำแนกผิดว่าเป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วไม่เป็นโรคเบาหวาน 1 คน และจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วเป็นโรคเบาหวาน 17 คน

4.1.2.3 การทำนายการเป็นโรคเบาหวาน

การทำนายการเป็นโรคเบาหวานจากข้อมูลทดสอบด้วยไฟล์ Diabetes-70.csv จำนวน 123 คน ได้ผลดังนี้

ตารางที่ 4.17 ผลในส่วนของการทำนายชุดข้อมูลทดสอบ ด้วยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ

ตัวอย่าง (instance)	ค่าที่แท้จริง (actual)	ค่าที่ทำนายได้ (predicted)	ความคลาดเคลื่อน (error)	กาทำนาย (prediction)
1	1:Negative	1:Negative		0.955
2	1:Negative	1:Negative		0.881
3	2:Positive	1:Negative	+	0.784
4	1:Negative	1:Negative		0.881
5	1:Negative	1:Negative		0.955
6	2:Positive	1:Negative	+	0.769
7	1:Negative	1:Negative		0.955
8	1:Negative	1:Negative		0.955
9	1:Negative	1:Negative		0.955
10	1:Negative	1:Negative		0.881
11	1:Negative	1:Negative		0.955
12	1:Negative	1:Negative		0.881
13	1:Negative	1:Negative		0.955
14	1:Negative	1:Negative		0.955
15	1:Negative	1:Negative		0.955
16	1:Negative	1:Negative		0.955
17	1:Negative	1:Negative		0.955
18	1:Negative	1:Negative		0.955

ตัวอย่าง (instance)	ค่าที่แท้จริง (actual)	ค่าที่ทำนายได้ (predicted)	ความคลาดเคลื่อน (error)	กาทำนาย (prediction)
19	1:Negative	1:Negative		0.955
20	1:Negative	1:Negative		0.955
21	1:Negative	1:Negative		0.955
22	1:Negative	1:Negative		0.955
23	1:Negative	1:Negative		0.955
24	1:Negative	1:Negative		0.955
25	1:Negative	1:Negative		0.955
26	1:Negative	1:Negative		0.881
27	1:Negative	1:Negative		0.955
28	1:Negative	1:Negative		0.881
29	1:Negative	1:Negative		0.955
30	1:Negative	1:Negative		0.955
31	1:Negative	1:Negative		0.881
32	1:Negative	1:Negative		0.955
33	1:Negative	1:Negative		0.955
34	1:Negative	1:Negative		0.881
35	1:Negative	1:Negative		0.955
36	1:Negative	1:Negative		0.955

ตารางที่ 4.17 ผลในส่วนของการทำนายชุดข้อมูลทดสอบ ด้วยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ (ต่อ)

ตัวอย่าง (instance)	ค่าที่แท้จริง (actual)	ค่าที่ทำนายได้ (predicted)	ความคลาดเคลื่อน (error)	การทำนาย (prediction)
37	1:Negative	1:Negative		0.881
38	1:Negative	1:Negative		0.955
39	1:Negative	1:Negative		0.955
40	1:Negative	1:Negative		0.955
41	1:Negative	1:Negative		0.955
42	1:Negative	1:Negative		0.955
43	1:Negative	1:Negative		0.955
44	1:Negative	1:Negative		0.881
45	1:Negative	1:Negative		0.955
46	1:Negative	1:Negative		0.955
47	2:Positive	1:Negative	+	0.769
48	1:Negative	1:Negative		0.955
49	1:Negative	1:Negative		0.955
50	1:Negative	1:Negative		0.894
51	1:Negative	1:Negative		0.955
52	1:Negative	1:Negative		0.955
53	1:Negative	1:Negative		0.955
54	1:Negative	1:Negative		0.955
55	1:Negative	1:Negative		0.881
56	1:Negative	1:Negative		0.955
57	1:Negative	1:Negative		0.955
58	1:Negative	1:Negative		0.955
59	1:Negative	1:Negative		1
60	1:Negative	1:Negative		0.955
61	2:Positive	1:Negative	+	1
62	1:Negative	1:Negative		0.955
63	1:Negative	1:Negative		0.955
64	1:Negative	1:Negative		0.955
65	2:Positive	1:Negative	+	0.769
66	1:Negative	1:Negative		0.955
67	1:Negative	1:Negative		0.881
68	1:Negative	1:Negative		0.955
69	2:Positive	1:Negative	+	0.955
70	1:Negative	1:Negative		0.881
71	1:Negative	1:Negative		0.837
72	1:Negative	1:Negative		0.837
73	1:Negative	1:Negative		0.955
74	2:Positive	1:Negative	+	0.955
75	1:Negative	1:Negative		0.955
76	1:Negative	1:Negative		0.955
77	1:Negative	1:Negative		0.881
78	1:Negative	1:Negative		0.955
79	1:Negative	1:Negative		0.955
80	1:Negative	1:Negative		0.955
81	1:Negative	1:Negative		0.955
82	2:Positive	1:Negative	+	0.667
83	1:Negative	1:Negative		0.955
84	1:Negative	1:Negative		0.955
85	1:Negative	1:Negative		0.955
86	1:Negative	1:Negative		0.955
87	1:Negative	1:Negative		0.955
88	1:Negative	1:Negative		0.955
89	1:Negative	1:Negative		0.955
90	1:Negative	1:Negative		0.955

ตารางที่ 4.17 ผลในส่วนของการทำนายชุดข้อมูลทดสอบ ด้วยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ (ต่อ)

ตัวอย่าง (instance)	ค่าที่แท้จริง (actual)	ค่าที่ทำนายได้ (predicted)	ความคลาดเคลื่อน (error)	การทำนาย (prediction)
91	1:Negative	1:Negative		0.955
92	1:Negative	1:Negative		0.955
93	1:Negative	1:Negative		0.881
94	1:Negative	1:Negative		0.955
95	1:Negative	1:Negative		0.881
96	1:Negative	1:Negative		0.648
97	1:Negative	1:Negative		0.955
98	1:Negative	1:Negative		0.955
99	1:Negative	1:Negative		0.955
100	1:Negative	1:Negative		0.955
101	1:Negative	1:Negative		0.955
102	1:Negative	1:Negative		0.955
103	1:Negative	1:Negative		0.955
104	1:Negative	1:Negative		0.955
105	1:Negative	1:Negative		0.955
106	1:Negative	1:Negative		0.955
107	1:Negative	1:Negative		0.881
108	1:Negative	1:Negative		0.955
109	1:Negative	1:Negative		0.881
110	1:Negative	1:Negative		0.955
111	1:Negative	1:Negative		0.955
112	1:Negative	1:Negative		0.881
113	1:Negative	1:Negative		0.881
114	1:Negative	1:Negative		0.955
115	1:Negative	1:Negative		0.955
116	1:Negative	1:Negative		0.955
117	1:Negative	1:Negative		0.955
118	1:Negative	1:Negative		0.955
119	1:Negative	1:Negative		0.955
120	1:Negative	1:Negative		0.955
121	1:Negative	1:Negative		0.955
122	1:Negative	1:Negative		0.955
123	1:Negative	1:Negative		0.837

จากตารางที่ 4.17 ค่าทำนายที่ทายผิดในช่อง error จะมีเครื่องหมาย + อยู่จำนวน 8 ค่า คือค่าที่ 3, 6, 47, 61, 65, 69, 74 และ 82 โดยที่ค่าที่ 3, 6, 47, 61, 65, 69, 74 และ 82 เป็นโรคเบาหวาน (class 2 : Positive) แต่ทำนายว่าไม่เป็นโรคเบาหวาน (class 1 : Negative)

ตารางที่ 4.18 ผลในส่วนของการสรุปผลจากการทำนายการเป็นโรคเบาหวาน

ด้วยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ

Correctly Classified Instances	115	93.4959 %
Incorrectly Classified Instances	8	6.5041%
Kappa statistic	0	
Mean absolute error	0.1137	
Root mean squared error	0.227	
Coverage of cases (0.95 level)	97.561%	
Total Number of Instances	123	

จากตารางที่ 4.18 โดยจากข้อมูลจำนวน 123 คน ทำนายข้อมูลถูกต้องจำนวน 115 คน คิดเป็น 93.4959% และทำนายข้อมูลไม่ถูกต้องจำนวน 8 คน คิดเป็น 6.5041% มีค่าสถิติแคปปา (Kappa statistic) คือ 0 แสดงว่าข้อมูลมีความสอดคล้องกันน้อยมาก มีค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) คือ 0.1137 ซึ่งมีค่าค่อนข้างน้อย แสดงว่าค่าที่ทำนายได้ใกล้เคียงกับค่าที่แท้จริงพอสมควร และมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) คือ $(0.227)^2 = 0.0515$ ซึ่งมีค่าน้อยมาก แสดงว่าตัวแบบมีความถูกต้องดี

ตารางที่ 4.19 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการทำนายการเป็นโรคเบาหวาน ด้วยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	1	1	0.935	1	0.966	Negative
	TN Rate	FN Rate	Precision	Recall	F-Measure	Class
	0	0	0	0	0	Positive
Weighted Avg.	0.935	0.935	0.874	0.935	0.904	

จากตารางที่ 4.19 สำหรับไม่คำตอบเป็นโรคเบาหวาน (Class = Negative) มีค่าอัตราความถูกต้องเชิงบวก = 1 ค่าอัตราความผิดพลาดเชิงบวก = 1 ค่าความแม่นยำ = 0.935 ค่าความระลึก = 1 และค่าความถ่วงดุล = 0.966 ส่วนคำตอบเป็นโรคเบาหวาน (Class = Positive) มีค่าอัตราความถูกต้องเชิงลบ = 0 ค่าอัตราความผิดพลาดเชิงลบ = 0 ค่าความแม่นยำ = 0 ค่าความระลึก = 0 และค่าความถ่วงดุล = 0

ตารางที่ 4.20 ผลในส่วนของเมตริกซ์ความสับสนจากการทำนายการเป็นโรคเบาหวาน ด้วยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ

		ผลการจำแนกเป็นโรคเบาหวาน	
		ไม่เป็น	เป็น
ค่าที่แท้จริง	ไม่เป็น	115	0
	เป็น	8	0

จากตารางที่ 4.20 มีข้อมูล 123 คน ตัวแบบสามารถทำนายข้อมูลได้ถูกต้อง 115 คน โดยมีจำนวนข้อมูลที่จำแนกถูกว่าไม่เป็นโรคเบาหวาน 115 คน และจำนวนข้อมูลที่จำแนกถูกว่าเป็นโรคเบาหวาน 0 คน ตัวแบบทำนายข้อมูลไม่ถูกต้อง 8 คน โดยมีจำนวนข้อมูลที่จำแนกผิดว่าเป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วไม่เป็นโรคเบาหวาน 0 คน และจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วเป็นโรคเบาหวาน 8 คน

4.1.3 วิธีโครงข่ายประสาท (Neural Network Method)

วิธีโครงข่ายประสาทใช้อัลกอริทึมชนิดเพอร์เซปตรอนแบบหลายชั้น (Multilayer Perception) ได้ผลการทำนายการเป็นโรคเบาหวานดังนี้

4.1.3.1 การสร้างตัวแบบการเป็นโรคเบาหวาน

สร้างตัวแบบการเป็นโรคเบาหวานจากชุดข้อมูลฝึกหัด ด้วยไฟล์ Diabetes-70.csv จำนวน 863 คน ได้ผลดังนี้

ตารางที่ 4.21 ผลในส่วนของการสรุปผลจากการสร้างตัวแบบการเป็นโรคเบาหวาน

ด้วยวิธีโครงข่ายประสาท

Correctly Classified Instances	770	89.2236 %
Incorrectly Classified Instances	93	10.7764 %
Kappa statistic	0.2532	
Mean absolute error	0.1195	
Root mean squared error	0.3151	
Relative absolute error	75.7592 %	
Root relative squared error	112.5264 %	
Coverage of cases (0.95 level)	92.3523 %	
Mean rel. region size (0.95 level)	55.9676 %	
Total Number of Instances	863	

จากตางรางที่ 4.21 จากข้อมูลจำนวน 863 คน ทำนายข้อมูลถูกต้องจำนวน 770 คน คิดเป็น 89.2236% และทำนายข้อมูลไม่ถูกต้องจำนวน 93 คน คิดเป็น 10.7764% มีค่าสถิติแคปปา (Kappa statistic) คือ 0.2532 แสดงว่าข้อมูลมีความสอดคล้องค่อนข้างน้อย มีค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) คือ 0.1195 ซึ่งมีค่าค่อนข้างน้อย แสดงว่าค่าที่ทำนายได้ใกล้เคียงกับค่าที่แท้จริงพอสมควร และมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) คือ $(0.3151)^2 = 0.0993$ ซึ่งมีค่าน้อยมาก แสดงว่าตัวแบบมีความถูกต้องดี

ตารางที่ 4.22 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการสร้างตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีโครงข่ายประสาท

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.949	0.716	0.934	0.949	0.942	Negative
	TN Rate	FN Rate	Precision	Recall	F-Measure	Class
	0.284	0.051	0.344	0.284	0.311	Positive
Weighted Avg.	0.892	0.659	0.883	0.892	0.887	

จากตารางที่ 4.22 สำหรับคำตอบไม่เป็นโรคเบาหวาน (Class = Negative) มีค่าอัตราความถูกต้องเชิงบวก = 0.949 ค่าอัตราความผิดพลาดเชิงบวก = 0.716 ค่าความแม่นยำ = 0.934 ค่าความระลึก = 0.949 และค่าความถ่วงดุล = 0.942 ส่วนคำตอบเป็นโรคเบาหวาน (Class = Positive) มีค่าอัตราความถูกต้องเชิงลบ = 0.284 ค่าอัตราความผิดพลาดเชิงลบ = 0.051 ค่าความแม่นยำ = 0.344 ค่าความระลึก = 0.284 และค่าความถ่วงดุล = 0.311

ตารางที่ 4.23 ผลในส่วนของเมตริกซ์ความสับสนจากการสร้างตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีโครงข่ายประสาท

		ผลการจำแนกเป็นโรคเบาหวาน	
		ไม่เป็น	เป็น
ค่าที่แท้จริง	ไม่เป็น	749	40
	เป็น	53	21

จากตารางที่ 4.23 มีข้อมูล 863 คน ตัวแบบสามารถทำนายข้อมูลได้ถูกต้อง 770 คน โดยมีจำนวนข้อมูลที่จำแนกถูกว่าไม่เป็นโรคเบาหวาน 749 คน และจำนวนข้อมูลที่จำแนกถูกว่าเป็นโรคเบาหวาน 21 คน ตัวแบบทำนายข้อมูลไม่ถูกต้อง 93 คน โดยมีจำนวนข้อมูลที่จำแนกผิดว่าเป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วไม่เป็นโรคเบาหวาน 40 คน และจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วเป็นโรคเบาหวาน 53 คน

4.1.3.2 การทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน

ทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวานจากชุดข้อมูลทดสอบประสิทธิภาพ จากไฟล์ Diabetes-20.csv จำนวน 247 คน ได้ผลดังนี้

ตารางที่ 4.24 ผลในส่วนของการสรุปผลจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีโครงข่ายประสาท

Correctly Classified Instances	218	88.2591 %
Incorrectly Classified Instances	29	11.7409 %
Kappa statistic	0.1083	
Mean absolute error	0.1217	
Root mean squared error	0.3141	
Coverage of cases (0.95 level)	93.5223 %	
Total Number of Instances	247	

จากตารางที่ 4.24 จากข้อมูลจำนวน 247 คน ทำนายข้อมูลถูกต้องจำนวน 218 คน คิดเป็น 88.2591% และทำนายข้อมูลไม่ถูกต้องจำนวน 29 คน คิดเป็น 11.7409% มีค่าสถิติแคปปา (Kappa statistic) คือ 0.1083 แสดงว่าข้อมูลมีความสอดคล้องกันค่อนข้างน้อย มีค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) คือ 0.1217 ซึ่งมีค่าค่อนข้างน้อย แสดงว่าค่าที่ทำนายได้ใกล้เคียงกับค่าที่แท้จริงพอสมควร และมีความค่าคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) คือ $(0.3141)^2 = 0.0987$ ซึ่งมีค่าน้อยมาก แสดงว่าตัวแบบมีความถูกต้องดี

ตารางที่ 4.25 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีโครงข่ายประสาท

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.935	0.824	0.939	0.935	0.937	Negative
	TN Rate	FN Rate	Precision	Recall	F-Measure	Class
	0.176	0.065	0.167	0.176	0.171	Positive
Weighted Avg.	0.883	0.771	0.886	0.883	0.884	

จากตารางที่ 4.25 สำหรับคำตอบไม่เป็นโรคเบาหวาน (Class = Negative) มีค่าอัตราความถูกต้องเชิงบวก = 0.935 ค่าอัตราความผิดพลาดเชิงบวก = 0.824 ค่าความแม่นยำ = 0.939 ค่าความระลึก = 0.935 และค่าความถ่วงดุล = 0.937 ส่วนคำตอบเป็นโรคเบาหวาน (Class = Positive) มีค่าอัตราความถูกต้องเชิงลบ = 0.176 ค่าอัตราความผิดพลาดเชิงลบ = 0.065 ค่าความแม่นยำ = 0.167 ค่าความระลึก = 0.176 และค่าความถ่วงดุล = 0.171

ตารางที่ 4.26 ผลในส่วนของเมตริกซ์ความสับสนจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีโครงข่ายประสาท

		ผลการจำแนกเป็นโรคเบาหวาน	
		ไม่เป็น	เป็น
ค่าที่แท้จริง	ไม่เป็น	215	15
	เป็น	14	3

จากตารางที่ 4.26 มีข้อมูล 247 คน ตัวแบบสามารถทำนายข้อมูลได้ถูกต้อง 218 คน โดยมีจำนวนข้อมูลที่จำแนกถูกว่าไม่เป็นโรคเบาหวาน 215 คน และจำนวนข้อมูลที่จำแนกถูกว่าเป็นโรคเบาหวาน 3 คน ตัวแบบทำนายข้อมูลไม่ถูกต้อง 29 คน โดยมีจำนวนข้อมูลที่จำแนกผิดว่าเป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วไม่เป็นโรคเบาหวาน 15 คน และจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วเป็นโรคเบาหวาน 14 คน

4.1.3.3 การทำนายการเป็นโรคเบาหวาน

การทำนายการเป็นโรคเบาหวานจากชุดข้อมูลทดสอบ ด้วยไฟล์ Diabetes-10.csv จำนวน 123 คน ได้ผลดังนี้

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

$$= \frac{123 \times 37}{37+2}$$

$$= 116.6923 \approx 117$$

จะได้จำนวนโหนดมากที่สุดในชั้นซ่อนประมาณ 117 โหนด โดยในที่นี้จะกำหนดจำนวนโหนดในชั้นซ่อนโดยแบ่งเป็นช่วงละ 12 คือ 12, 24, 36, 48, 60, 72, 84, 96, 108 และ 120 กำหนดค่าอัตราการเรียนรู้เป็น 0.1, 0.2, 0.3, 0.4 และ 0.5 ค่าโมเมนตัมเป็น 0.5, 0.6, 0.7, 0.8 และ 0.9 และใช้จำนวนรอบการฝึกหัดคือ 20,000 รอบ

ตารางที่ 4.27 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) ของการทำนายการเป็นโรคเบาหวาน ด้วยวิธีโครงข่ายประสาท สำหรับอัตราการเรียนรู้ η = 0.1 และโมเมนตัม α = 0.5, 0.6, 0.7, 0.8 และ 0.9

โครงข่าย	โมเมนตัม				
	0.5	0.6	0.7	0.8	0.9
19-12-2	0.0910	0.1067	0.0916	0.0980	0.0669
19-24-2	0.1001	0.0893	0.0879	0.0612	0.0708
19-36-2	0.1075	0.0701	0.0769	0.0786	0.0661
19-48-2	0.0765	0.0897	0.0858	0.0771	0.0751
19-60-2	0.0848	0.0884	0.0816	0.0844	0.0825
19-72-2	0.0830	0.0881	0.0870	0.0776	0.0757
19-84-2	0.0875	0.1071	0.0843	0.0973	0.0807
19-96-2	0.0890	0.0879	0.0795	0.0778	0.0933
19-108-2	0.0682	0.0713	0.0662	0.1099	0.0763
19-120-2	0.0787	0.0766	0.1019	0.1008	0.0968

จากตารางที่ 4.27 เมื่อกำหนดค่าอัตราการเรียนรู้เท่ากับ 0.1 และค่าโมเมนตัมเท่ากับ 0.8 จะพบว่าโครงข่ายประสาทที่ให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด (MSE = 0.0612) คือโครงข่ายประสาท 19-24-2 ซึ่งเป็นโครงข่ายประสาทที่มีโหนดข้อมูลเข้าเท่ากับ 19 โหนด โหนดซ่อนเท่ากับ 24 โหนด และโหนดข้อมูลออกเท่ากับ 2 โหนด

ตารางที่ 4.28 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) ของการทำนายการเป็นโรคเบาหวาน ด้วยวิธีโครงข่ายประสาท สำหรับอัตราการเรียนรู้ η = 0.2 และโมเมนตัม α = 0.5, 0.6, 0.7, 0.8 และ 0.9

โครงข่าย	โมเมนตัม				
	0.5	0.6	0.7	0.8	0.9
19-12-2	0.0821	0.1236	0.0970	0.0727	0.0645
19-24-2	0.0853	0.0809	0.0743	0.0899	0.0781
19-36-2	0.0954	0.0873	0.0597	0.0818	0.0699
19-48-2	0.1077	0.0773	0.0752	0.0738	0.0622
19-60-2	0.0946	0.0858	0.0913	0.0820	0.0652
19-72-2	0.0831	0.0819	0.0737	0.0776	0.0573
19-84-2	0.1044	0.0938	0.1032	0.0813	0.0726
19-96-2	0.0949	0.0800	0.0830	0.0864	0.0907
19-108-2	0.0811	0.0927	0.0726	0.0955	0.0787
19-120-2	0.0804	0.0965	0.0842	0.0775	0.0808

จากตารางที่ 4.28 เมื่อกำหนดค่าอัตราการเรียนรู้เท่ากับ 0.2 และค่าโมเมนตัมเท่ากับ 0.7 จะพบว่าโครงข่ายประสาทที่ให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด (MSE = 0.0597) คือโครงข่ายประสาท 19-36-2 ซึ่งเป็นโครงข่ายประสาทที่มีโหนดข้อมูลเข้าเท่ากับ 19 โหนด โหนดซ่อนเท่ากับ 36 โหนด และโหนดข้อมูลออกเท่ากับ 2 โหนด

ตารางที่ 4.29 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) ของการทำนายการเป็นโรคเบาหวาน ด้วยวิธีโครงข่ายประสาท สำหรับอัตราการเรียนรู้ η = 0.3 และโมเมนตัม α = 0.5, 0.6, 0.7, 0.8 และ 0.9

โครงข่าย	โมเมนตัม				
	0.5	0.6	0.7	0.8	0.9
19-12-2	0.1162	0.0911	0.0492	0.0643	0.0627
19-24-2	0.0396	0.0495	0.0765	0.0771	0.0817
19-36-2	0.0845	0.0903	0.0955	0.0712	0.0649
19-48-2	0.0992	0.0896	0.0746	0.0675	0.0812
19-60-2	0.0925	0.0767	0.0713	0.0860	0.0866
19-72-2	0.0870	0.0960	0.0562	0.0722	0.0594
19-84-2	0.0765	0.0916	0.0692	0.0663	0.0836
19-96-2	0.0879	0.0910	0.0804	0.0825	0.0812
19-108-2	0.0698	0.1108	0.1011	0.0852	0.0672
19-120-2	0.0988	0.0820	0.0793	0.0740	0.0795

จากตารางที่ 4.29 เมื่อกำหนดค่าอัตราการเรียนรู้เท่ากับ 0.3 และค่าโมเมนตัมเท่ากับ 0.5 จะพบว่าโครงข่ายประสาทที่ให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด (MSE = 0.0396) คือโครงข่ายประสาท 19-24-2 ซึ่งเป็นโครงข่ายประสาทที่มีโหนดข้อมูลเข้าเท่ากับ 19 โหนด โหนดซ่อนเท่ากับ 24 โหนด และโหนดข้อมูลออกเท่ากับ 2 โหนด

ตารางที่ 4.30 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) ของการทำนายการเป็นโรคเบาหวาน ด้วยวิธีโครงข่ายประสาท สำหรับอัตราการเรียนรู้ η = 0.4 และโมเมนตัม α = 0.5, 0.6, 0.7 0.8 และ 0.9

โครงข่าย	โมเมนตัม				
	0.5	0.6	0.7	0.8	0.9
19-12-2	0.1151	0.0609	0.0649	0.0846	0.0653
19-24-2	0.0846	0.0870	0.0931	0.0650	0.0615
19-36-2	0.0857	0.0658	0.0722	0.0914	0.0732
19-48-2	0.0880	0.0576	0.0728	0.0804	0.0589
19-60-2	0.0911	0.0777	0.0662	0.0781	0.0813
19-72-2	0.0908	0.0897	0.0944	0.0703	0.0591
19-84-2	0.1110	0.0942	0.0740	0.0726	0.0726
19-96-2	0.0895	0.0937	0.0692	0.0955	0.0645
19-108-2	0.0760	0.0886	0.0871	0.0604	0.0797
19-120-2	0.0817	0.0765	0.0651	0.1132	0.0590

จากตารางที่ 4.30 เมื่อกำหนดค่าอัตราการเรียนรู้เท่ากับ 0.4 และค่าโมเมนตัมเท่ากับ 0.9 จะพบว่าโครงข่ายประสาทที่ให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด (MSE = 0.0589) คือโครงข่ายประสาท 19-48-2 ซึ่งเป็นโครงข่ายประสาทที่มีโหนดข้อมูลเข้าเท่ากับ 19 โหนด โหนดซ่อนเท่ากับ 48 โหนด และโหนดข้อมูลออกเท่ากับ 2 โหนด

ตารางที่ 4.31 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) ของการทำนายการเป็นโรคเบาหวาน ด้วยวิธีโครงข่ายประสาท สำหรับอัตราการเรียนรู้ η = 0.5 และโมเมนตัม α = 0.5, 0.6, 0.7 0.8 และ 0.9

โครงข่าย	โมเมนตัม				
	0.5	0.6	0.7	0.8	0.9
19-12-2	0.0853	0.0529	0.0836	0.0680	0.0568
19-24-2	0.0959	0.0679	0.0710	0.0829	0.0647
19-36-2	0.0618	0.0843	0.0771	0.0645	0.0650
19-48-2	0.0809	0.0879	0.0662	0.0654	0.0673
19-60-2	0.0729	0.0861	0.0732	0.0739	0.0650
19-72-2	0.0825	0.0953	0.0792	0.0730	0.0731
19-84-2	0.0776	0.0785	0.0948	0.0652	0.0650
19-96-2	0.1000	0.0657	0.0924	0.0751	0.0650
19-108-2	0.0776	0.0840	0.0736	0.0894	0.0580
19-120-2	0.1058	0.0970	0.0746	0.0788	0.0944

จากตารางที่ 4.31 เมื่อกำหนดค่าอัตราการเรียนรู้เท่ากับ 0.5 และค่าโมเมนตัมเท่ากับ 0.6 จะพบว่าโครงข่ายประสาทที่ให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด (MSE = 0.0529) คือโครงข่ายประสาท 19-12-2 ซึ่งเป็นโครงข่ายประสาทที่มีโหนดข้อมูลเข้าเท่ากับ 19 โหนด โหนดซ่อนเท่ากับ 12 โหนด และโหนดข้อมูลออกเท่ากับ 2 โหนด

จากการเปรียบเทียบทั้งหมด เมื่อกำหนดค่าอัตราการเรียนรู้เท่ากับ 0.3 และค่าโมเมนตัมเท่ากับ 0.5 จะพบว่าโครงข่ายประสาทที่ให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด (MSE = 0.0396) คือโครงข่ายประสาท 19-24-2 ซึ่งเป็นโครงข่ายประสาทที่มีโหนดข้อมูลเข้าเท่ากับ 19 โหนด โหนดซ่อนเท่ากับ 24 โหนด และโหนดข้อมูลออกเท่ากับ 2 โหนด ได้ผลดังนี้

ตารางที่ 4.32 ผลในส่วนของการทำนายชุดข้อมูลทดสอบ ด้วยวิธีโครงข่ายประสาท

ตัวอย่าง (instance)	ค่าที่แท้จริง (actual)	ค่าที่ทำนายได้ (predicted)	ความคลาดเคลื่อน (error)	การทำนาย (prediction)
1	1:Negative	1:Negative		1
2	1:Negative	1:Negative		1
3	2:Positive	1:Negative	+	1
4	1:Negative	1:Negative		1
5	1:Negative	1:Negative		1
6	2:Positive	2:Positive		0.999
7	1:Negative	1:Negative		1
8	1:Negative	1:Negative		1
9	1:Negative	1:Negative		1
10	1:Negative	1:Negative		0.715
11	1:Negative	2:Positive	+	0.906
12	1:Negative	1:Negative		1
13	1:Negative	1:Negative		1
14	1:Negative	1:Negative		1
15	1:Negative	2:Positive	+	0.835
16	1:Negative	1:Negative		1
17	1:Negative	1:Negative		1
18	1:Negative	1:Negative		1
19	1:Negative	1:Negative		1
20	1:Negative	1:Negative		1
21	1:Negative	1:Negative		1
22	1:Negative	1:Negative		1
23	1:Negative	1:Negative		1
24	1:Negative	1:Negative		1
25	1:Negative	1:Negative		1
26	1:Negative	1:Negative		1
27	1:Negative	1:Negative		1

ตัวอย่าง (instance)	ค่าที่แท้จริง (actual)	ค่าที่ทำนายได้ (predicted)	ความคลาดเคลื่อน (error)	การทำนาย (prediction)
28	1:Negative	1:Negative		1
29	1:Negative	1:Negative		1
30	1:Negative	1:Negative		0.571
31	1:Negative	1:Negative		1
32	1:Negative	1:Negative		1
33	1:Negative	1:Negative		1
34	1:Negative	1:Negative		1
35	1:Negative	1:Negative		1
36	1:Negative	1:Negative		1
37	1:Negative	1:Negative		1
38	1:Negative	1:Negative		1
39	1:Negative	1:Negative		1
40	1:Negative	1:Negative		1
41	1:Negative	1:Negative		1
42	1:Negative	1:Negative		1
43	1:Negative	1:Negative		1
44	1:Negative	1:Negative		1
45	1:Negative	1:Negative		1
46	1:Negative	1:Negative		1
47	2:Positive	2:Positive		0.759
48	1:Negative	1:Negative		1
49	1:Negative	1:Negative		1
50	1:Negative	1:Negative		0.999
51	1:Negative	1:Negative		1
52	1:Negative	1:Negative		1
53	1:Negative	1:Negative		1
54	1:Negative	1:Negative		1

ตารางที่ 4.32 ผลในส่วนของการทำนายชุดข้อมูลทดสอบ ด้วยวิธีโครงข่ายประสาท (ต่อ)

ตัวอย่าง (instance)	ค่าที่แท้จริง (actual)	ค่าที่ทำนายได้ (predicted)	ความคลาดเคลื่อน (error)	การทำนาย (prediction)
55	1:Negative	1:Negative		1
56	1:Negative	1:Negative		1
57	1:Negative	1:Negative		1
58	1:Negative	1:Negative		1
59	1:Negative	1:Negative		1
60	1:Negative	1:Negative		1
61	2:Positive	1:Negative	+	1
62	1:Negative	1:Negative		1
63	1:Negative	1:Negative		1
64	1:Negative	1:Negative		1
65	2:Positive	2:Positive		0.884
66	1:Negative	1:Negative		1
67	1:Negative	1:Negative		1
68	1:Negative	1:Negative		1
69	2:Positive	1:Negative	+	1
70	1:Negative	1:Negative		1
71	1:Negative	1:Negative		1
72	1:Negative	1:Negative		1
73	1:Negative	1:Negative		1
74	2:Positive	2:Positive		0.934
75	1:Negative	1:Negative		1
76	1:Negative	1:Negative		1
77	1:Negative	1:Negative		1
78	1:Negative	1:Negative		1
79	1:Negative	1:Negative		1
80	1:Negative	1:Negative		1
81	1:Negative	1:Negative		1

ตัวอย่าง (instance)	ค่าที่แท้จริง (actual)	ค่าที่ทำนายได้ (predicted)	ความคลาดเคลื่อน (error)	การทำนาย (prediction)
82	2:Positive	2:Positive		0.976
83	1:Negative	1:Negative		1
84	1:Negative	1:Negative		1
85	1:Negative	1:Negative		1
86	1:Negative	1:Negative		1
87	1:Negative	1:Negative		1
88	1:Negative	1:Negative		1
89	1:Negative	1:Negative		1
90	1:Negative	1:Negative		1
91	1:Negative	1:Negative		1
92	1:Negative	1:Negative		1
93	1:Negative	1:Negative		1
94	1:Negative	1:Negative		1
95	1:Negative	1:Negative		0.966
96	1:Negative	1:Negative		0.999
97	1:Negative	1:Negative		0.999
98	1:Negative	1:Negative		1
99	1:Negative	1:Negative		1
100	1:Negative	1:Negative		1
101	1:Negative	1:Negative		1
102	1:Negative	1:Negative		0.959
103	1:Negative	1:Negative		1
104	1:Negative	1:Negative		1
105	1:Negative	1:Negative		1
106	1:Negative	1:Negative		1
107	1:Negative	1:Negative		1
108	1:Negative	1:Negative		1

ตารางที่ 4.32 ผลในส่วนของการทำนายชุดข้อมูลทดสอบ ด้วยวิธีโครงข่ายประสาท (ต่อ)

ตัวอย่าง (instance)	ค่าที่แท้จริง (actual)	ค่าที่ทำนายได้ (predicted)	ความคลาดเคลื่อน (error)	การทำนาย (prediction)
109	1:Negative	1:Negative		1
110	1:Negative	1:Negative		1
111	1:Negative	1:Negative		1
112	1:Negative	1:Negative		1
113	1:Negative	1:Negative		1
114	1:Negative	1:Negative		1
115	1:Negative	1:Negative		1
116	1:Negative	1:Negative		1
117	1:Negative	1:Negative		1
118	1:Negative	1:Negative		1
119	1:Negative	1:Negative		1
120	1:Negative	1:Negative		1
121	1:Negative	1:Negative		1
122	1:Negative	1:Negative		1
123	1:Negative	1:Negative		0.946

จากตารางที่ 4.32 ค่าทำนายที่ทายผิดในช่อง error จะมีเครื่องหมาย + อยู่จำนวน 5 ค่า คือค่าที่ 3, 11, 15, 61 และ 69 โดยที่ค่าที่ 3, 61 และ 69 เป็นโรคเบาหวาน (class 2 : Positive) แต่ทำนายว่าไม่เป็นโรคเบาหวาน (class 1 : Negative) และค่าที่ 11 และ 15 ไม่เป็นโรคเบาหวาน แต่ทำนายว่าเป็นโรคเบาหวาน

ตารางที่ 4.33 ผลในส่วนของการสรุปผลจากการทำนายการเป็นโรคเบาหวาน ด้วยวิธีโครงข่ายประสาท

Correctly Classified Instances	118	95.935 %
Incorrectly Classified Instances	5	4.065 %
Kappa statistic	0.6451	
Mean absolute error	0.0491	
Root mean squared error	0.1989	
Relative absolute error	34.9308 %	
Root relative squared error	80.3311 %	
Coverage of cases (0.95 level)	97.5610 %	
Mean rel. region size (0.95 level)	53.2520 %	
Total Number of Instances	123	

จากตารางที่ 4.33 จากข้อมูลจำนวน 123 คน ทำนายข้อมูลถูกต้องจำนวน 118 คน คิดเป็น 95.935% และทำนายข้อมูลไม่ถูกต้องจำนวน 5 คน คิดเป็น 4.065% มีค่าสถิติแคปปา (Kappa statistic) คือ 0.6451 แสดงว่าข้อมูลมีความสอดคล้องกันปานกลาง มีค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) คือ 0.0491 ซึ่งมีค่าน้อยมาก แสดงว่าค่าที่ทำนายได้ใกล้เคียงกับค่าที่แท้จริง และมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) คือ $(0.1989)^2 = 0.0396$ ซึ่งมีค่าน้อยมาก แสดงว่าตัวแบบมีความถูกต้องดี

ตารางที่ 4.34 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการทำนายการเป็นโรคเบาหวาน ด้วยวิธีโครงข่ายประสาท

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.983	0.375	0.974	0.983	0.978	Negative
	TN Rate	FN Rate	Precision	Recall	F-Measure	Class
	0.625	0.017	0.714	0.625	0.667	Positive
Weighted Avg.	0.959	0.352	0.957	0.959	0.958	

จากตารางที่ 4.34 สำหรับคำตอบไม่เป็นโรคเบาหวาน (Class = Negative) มีค่าอัตราความถูกต้องเชิงบวก = 0.983 ค่าอัตราความผิดพลาดเชิงบวก = 0.375 ค่าความแม่นยำ = 0.974 ค่าความระลึก = 0.983 และค่าความถ่วงดุล = 0.978 ส่วนคำตอบเป็นโรคเบาหวาน (Class = Positive) มีค่าอัตราความถูกต้องเชิงลบ = 0.625 ค่าอัตราความผิดพลาดเชิงลบ = 0.017 ค่าความแม่นยำ = 0.714 ค่าความระลึก = 0.625 และค่าความถ่วงดุล = 0.667

ตารางที่ 4.35 ผลในส่วนของเมตริกซ์ความสับสนจากการทำนายการเป็นโรคเบาหวาน ด้วยวิธีโครงข่ายประสาท

		ผลการจำแนกเป็นโรคเบาหวาน	
		ไม่เป็น	เป็น
ค่าที่แท้จริง	ไม่เป็น	113	2
	เป็น	3	5

จากตารางที่ 4.35 มีข้อมูล 123 คน ตัวแบบสามารถทำนายข้อมูลได้ถูกต้อง 118 คน โดยมีจำนวนข้อมูลที่จำแนกถูกว่าไม่เป็นโรคเบาหวาน 113 คน และจำนวนข้อมูลที่จำแนกถูกว่าเป็นโรคเบาหวาน 5 คน ตัวแบบทำนายข้อมูลไม่ถูกต้อง 5 คน โดยมีจำนวนข้อมูลที่จำแนกผิดว่าเป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วไม่เป็นโรคเบาหวาน 2 คน และจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วเป็นโรคเบาหวาน 3 คน

4.1.4 วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine Method)

วิธีซัพพอร์ตเวกเตอร์แมชชีนนั้นมีฟังก์ชันหลายฟังก์ชัน แต่ในงานวิจัยนี้ใช้อัลกอริทึม SMO ชนิดโพลิโนเมียลเคอร์เนล (Polynomial Kernel) ในการทำนายการเป็นโรคเบาหวาน ได้ผลดังนี้

4.1.4.1 การสร้างตัวแบบการเป็นโรคเบาหวาน

สร้างตัวแบบการเป็นโรคเบาหวานจากชุดข้อมูลฝึกหัด ด้วยไฟล์ Diabetes-70.csv จำนวน 863 คน ได้ผลดังนี้

ตารางที่ 4.36 ผลในส่วนของการสรุปผล จากการสร้างตัวแบบการเป็นโรคเบาหวาน

ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน

Correctly Classified Instances	786	91.0776 %
Incorrectly Classified Instances	77	8.9224 %
Kappa statistic	-0.0067	
Mean absolute error	0.0892	
Root mean squared error	0.2987	
Relative absolute error	56.5870 %	
Root relative squared error	106.6780 %	
Coverage of cases (0.95 level)	91.0776 %	
Mean rel. region size (0.95 level)	50 %	
Total Number of Instances	863	

จากตารางที่ 4.36 จากข้อมูลจำนวน 863 คน ทำนายข้อมูลถูกต้องจำนวน 786 คน คิดเป็น 91.0776% และทำนายข้อมูลไม่ถูกต้องจำนวน 77 คน คิดเป็น 8.9224% มีค่าสถิติแคปปา (Kappa statistic) คือ -0.0067 แสดงว่าข้อมูลมีความสอดคล้องกันน้อยมาก มีค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) คือ 0.0892 ซึ่งมีค่าน้อยมาก แสดงว่าค่าที่ทำนายได้ใกล้เคียงกับค่าที่แท้จริง และมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) คือ $(0.2987)^2 = 0.0892$ ซึ่งมีค่าน้อยมาก แสดงว่าตัวแบบมีความถูกต้องดี

ตารางที่ 4.37 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการสร้างตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.996	1	0.914	0.996	0.953	Negative
	TN Rate	FN Rate	Precision	Recall	F-Measure	Class
	0	0.004	0	0	0	Positive
Weighted Avg.	0.911	0.915	0.836	0.911	0.872	

จากตารางที่ 4. 37 สำหรับคำตอบไม่เป็นโรคเบาหวาน (Class = Negative) มีค่าอัตราความถูกต้องเชิงบวก = 0.996 ค่าอัตราความผิดพลาดเชิงบวก = 1 ค่าความแม่นยำ = 0.914 ค่าความระลึก = 0.996 และค่าความถ่วงดุล = 0.953 ส่วนคำตอบเป็นโรคเบาหวาน (Class = Positive) มีค่าอัตราความถูกต้องเชิงลบ = 0 ค่าอัตราความผิดพลาดเชิงลบ = 0.004 ค่าความแม่นยำ = 0 ค่าความระลึก = 0 และค่าความถ่วงดุล = 0

ตารางที่ 4.38 ผลในส่วนของเมตริกซ์ความสับสนจากการสร้างตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน

		ผลการจำแนกเป็นโรคเบาหวาน	
		ไม่เป็น	เป็น
ค่าที่แท้จริง	ไม่เป็น	783	3
	เป็น	74	0

จากตารางที่ 4.38 มีข้อมูล 863 คน ตัวแบบสามารถทำนายข้อมูลได้ถูกต้อง 783 คน โดยมีจำนวนข้อมูลที่จำแนกถูกว่าไม่เป็นโรคเบาหวาน 783 คน และจำนวนข้อมูลที่จำแนกถูกว่าเป็นโรคเบาหวาน 0 คน ตัวแบบทำนายข้อมูลไม่ถูกต้อง 77 คน โดยมีจำนวนข้อมูลที่จำแนกผิดว่าเป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วไม่เป็นโรคเบาหวาน 3 คน และจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วเป็นโรคเบาหวาน 74 คน

4.1.4.2 การทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน

ทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวานจากชุดข้อมูลทดสอบประสิทธิภาพ ด้วยไฟล์ Diabetes-20.csv จำนวน 247 คน ได้ผลดังนี้

ตารางที่ 4.39 ผลในส่วนของการสรุปผลจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน

Correctly Classified Instances	230	93.1174 %
Incorrectly Classified Instances	17	6.8826 %
Kappa statistic	0	
Mean absolute error	0.0688	
Root mean squared error	0.2623	
Coverage of cases (0.95 level)	93.1174 %	
Total Number of Instances	247	

จากตารางที่ 4.39 จากข้อมูลจำนวน 247 คน ทำนายข้อมูลถูกต้องจำนวน 230 คน คิดเป็น 93.1174% และทำนายข้อมูลไม่ถูกต้องจำนวน 17 คน คิดเป็น 6.8826% มีค่าสถิติแคปปา (Kappa statistic) คือ 0 แสดงว่าข้อมูลมีความสอดคล้องกันน้อยมาก มีค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) คือ 0.0688 ซึ่งมีค่าน้อยมาก แสดงว่าค่าที่ทำนายได้ใกล้เคียงกับค่าที่แท้จริง และมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) คือ $(0.2623)^2 = 0.0688$ ซึ่งมีค่าน้อยมาก แสดงว่าตัวแบบมีความถูกต้องดี

ตารางที่ 4.40 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	1	1	0.931	1	0.964	Negative
	TN Rate	FN Rate	Precision	Recall	F-Measure	Class
	0	0	0	0	0	Positive
Weighted Avg.	0.931	0.931	0.867	0.931	0.898	

จากตารางที่ 4. 40 สำหรับคำตอบไม่เป็นโรคเบาหวาน (Class = Negative) มีค่าอัตราความถูกต้องเชิงบวก = 1 ค่าอัตราความผิดพลาดเชิงบวก = 1 ค่าความแม่นยำ = 0.931 ค่าความระลึก = 1 และค่าความถ่วงดุล = 0.964 ส่วนคำตอบเป็นโรคเบาหวาน (Class = Positive) มีค่าอัตราความถูกต้องเชิงลบ = 0 ค่าอัตราความผิดพลาดเชิงลบ = 0 ค่าความแม่นยำ = 0 ค่าความระลึก = 0 และค่าความถ่วงดุล = 0

ตารางที่ 4.41 ผลในส่วนของเมตริกซ์ความสับสนจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน

		ผลการจำแนกเป็นโรคเบาหวาน	
		ไม่เป็น	เป็น
ค่าที่แท้จริง	ไม่เป็น	230	0
	เป็น	17	0

จากตารางที่ 4.41 มีข้อมูล 247 คน ตัวแบบสามารถทำนายข้อมูลได้ถูกต้อง 230 คน โดยมีจำนวนข้อมูลที่จำแนกถูกว่าไม่เป็นโรคเบาหวาน 230 คน และจำนวนข้อมูลที่จำแนกถูกว่าเป็นโรคเบาหวาน 0 คน ตัวแบบทำนายข้อมูลไม่ถูกต้อง 17 คน โดยมีจำนวนข้อมูลที่จำแนกผิดว่าเป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วไม่เป็นโรคเบาหวาน 0 คน และจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วเป็นโรคเบาหวาน 17 คน

4.1.4.3 การทำนายการเป็นโรคเบาหวาน

การทำนายการเป็นโรคเบาหวานจากชุดข้อมูลทดสอบ ด้วยไฟล์ Diabetes-10.csv จำนวน 123 คน ได้ผลดังนี้

ตารางที่ 4.42 ผลในส่วนของการทำนายชุดข้อมูลทดสอบ ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน

ตัวอย่าง (instance)	ค่าที่แท้จริง (actual)	ค่าที่ทำนายได้ (predicted)	ความคลาดเคลื่อน (error)	การทำนาย (prediction)
1	1:Negative	1:Negative		1
2	1:Negative	1:Negative		1
3	2:Positive	1:Negative	+	1
4	1:Negative	1:Negative		1
5	1:Negative	1:Negative		1
6	2:Positive	1:Negative	+	1
7	1:Negative	1:Negative		1
8	1:Negative	1:Negative		1
9	1:Negative	1:Negative		1
10	1:Negative	1:Negative		1
11	1:Negative	1:Negative		1
12	1:Negative	1:Negative		1
13	1:Negative	1:Negative		1
14	1:Negative	1:Negative		1
15	1:Negative	1:Negative		1
16	1:Negative	1:Negative		1

ตัวอย่าง (instance)	ค่าที่แท้จริง (actual)	ค่าที่ทำนายได้ (predicted)	ความคลาดเคลื่อน (error)	การทำนาย (prediction)
17	1:Negative	1:Negative		1
18	1:Negative	1:Negative		1
19	1:Negative	1:Negative		1
20	1:Negative	1:Negative		1
21	1:Negative	1:Negative		1
22	1:Negative	1:Negative		1
23	1:Negative	1:Negative		1
24	1:Negative	1:Negative		1
25	1:Negative	1:Negative		1
26	1:Negative	1:Negative		1
27	1:Negative	1:Negative		1
28	1:Negative	1:Negative		1
29	1:Negative	1:Negative		1
30	1:Negative	1:Negative		1
31	1:Negative	1:Negative		1
32	1:Negative	1:Negative		1

ตารางที่ 4.42 ผลในส่วนของการทำนายชุดข้อมูลทดสอบ ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน (ต่อ)

ตัวอย่าง (instance)	ค่าที่แท้จริง (actual)	ค่าที่ทำนายได้ (predicted)	ความคลาดเคลื่อน (error)	การทำนาย (prediction)
33	1:Negative	1:Negative		1
34	1:Negative	1:Negative		1
35	1:Negative	1:Negative		1
36	1:Negative	1:Negative		1
37	1:Negative	1:Negative		1
38	1:Negative	1:Negative		1
39	1:Negative	1:Negative		1
40	1:Negative	1:Negative		1
41	1:Negative	1:Negative		1
42	1:Negative	1:Negative		1
43	1:Negative	1:Negative		1
44	1:Negative	1:Negative		1
45	1:Negative	1:Negative		1
46	1:Negative	1:Negative		1
47	2:Positive	1:Negative	+	1
48	1:Negative	1:Negative		1
49	1:Negative	1:Negative		1
50	1:Negative	1:Negative		1
51	1:Negative	1:Negative		1
52	1:Negative	1:Negative		1
53	1:Negative	1:Negative		1
54	1:Negative	1:Negative		1
55	1:Negative	1:Negative		1
56	1:Negative	1:Negative		1
57	1:Negative	1:Negative		1
58	1:Negative	1:Negative		1
59	1:Negative	1:Negative		1

ตัวอย่าง (instance)	ค่าที่แท้จริง (actual)	ค่าที่ทำนายได้ (predicted)	ความคลาดเคลื่อน (error)	การทำนาย (prediction)
60	1:Negative	1:Negative		1
61	2:Positive	1:Negative	+	1
62	1:Negative	1:Negative		1
63	1:Negative	1:Negative		1
64	1:Negative	1:Negative		1
65	2:Positive	1:Negative	+	1
66	1:Negative	1:Negative		1
67	1:Negative	1:Negative		1
68	1:Negative	1:Negative		1
69	2:Positive	1:Negative	+	1
70	1:Negative	1:Negative		1
71	1:Negative	1:Negative		1
72	1:Negative	1:Negative		1
73	1:Negative	1:Negative		1
74	2:Positive	1:Negative	+	1
75	1:Negative	1:Negative		1
76	1:Negative	1:Negative		1
77	1:Negative	1:Negative		1
78	1:Negative	1:Negative		1
79	1:Negative	1:Negative		1
80	1:Negative	1:Negative		1
81	1:Negative	1:Negative		1
82	2:Positive	1:Negative	+	1
83	1:Negative	1:Negative		1
84	1:Negative	1:Negative		1
85	1:Negative	1:Negative		1
86	1:Negative	1:Negative		1

ตารางที่ 4.42 ผลในส่วนของการทำนายชุดข้อมูลทดสอบ ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน (ต่อ)

ตัวอย่าง (instance)	ค่าที่แท้จริง (actual)	ค่าที่ทำนายได้ (predicted)	ความคลาดเคลื่อน (error)	การทำนาย (prediction)
87	1:Negative	1:Negative		1
88	1:Negative	1:Negative		1
89	1:Negative	1:Negative		1
90	1:Negative	1:Negative		1
91	1:Negative	1:Negative		1
92	1:Negative	1:Negative		1
93	1:Negative	1:Negative		1
94	1:Negative	1:Negative		1
95	1:Negative	1:Negative		1
96	1:Negative	1:Negative		1
97	1:Negative	1:Negative		1
98	1:Negative	1:Negative		1
99	1:Negative	1:Negative		1
100	1:Negative	1:Negative		1
101	1:Negative	1:Negative		1
102	1:Negative	1:Negative		1
103	1:Negative	1:Negative		1
104	1:Negative	1:Negative		1
105	1:Negative	1:Negative		1

ตัวอย่าง (instance)	ค่าที่แท้จริง (actual)	ค่าที่ทำนายได้ (predicted)	ความคลาดเคลื่อน (error)	การทำนาย (prediction)
106	1:Negative	1:Negative		1
107	1:Negative	1:Negative		1
108	1:Negative	1:Negative		1
109	1:Negative	1:Negative		1
110	1:Negative	1:Negative		1
111	1:Negative	1:Negative		1
112	1:Negative	1:Negative		1
113	1:Negative	1:Negative		1
114	1:Negative	1:Negative		1
115	1:Negative	1:Negative		1
116	1:Negative	1:Negative		1
117	1:Negative	1:Negative		1
118	1:Negative	1:Negative		1
119	1:Negative	1:Negative		1
120	1:Negative	1:Negative		1
121	1:Negative	1:Negative		1
122	1:Negative	1:Negative		1
123	1:Negative	1:Negative		1

จากตารางที่ 4.42 ค่าทำนายที่ทายผิดในช่อง error จะมีเครื่องหมาย + อยู่จำนวน 8 ค่า คือค่าที่ 3, 6, 47, 61, 65, 69, 74 และ 82 โดยที่ค่าที่ 3, 6, 47, 61, 65, 69, 74 และ 82 เป็นโรคเบาหวาน (class 2 : Positive) แต่ทำนายว่าไม่เป็นโรคเบาหวาน (class 1 : Negative)

ตารางที่ 4.43 ผลในส่วนของการสรุปผลจากการทำนายการเป็นโรคเบาหวาน ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน

Correctly Classified Instances	115	93.4959 %
Incorrectly Classified Instances	8	6.5041 %
Kappa statistic	0	
Mean absolute error	0.0650	
Root mean squared error	0.2550	
Coverage of cases (0.95 level)	93.4959 %	
Total Number of Instances	123	

จากตารางที่ 4.43 โดยจากข้อมูลจำนวน 123 คน ทำนายข้อมูลถูกต้องจำนวน 115 คน คิดเป็น 93.4959% และทำนายข้อมูลไม่ถูกต้องจำนวน 8 คน คิดเป็น 6.5041% มีค่าสถิติแคปปา (Kappa statistic) คือ 0 แสดงว่าข้อมูลมีความสอดคล้องกันน้อยมาก มีค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) คือ 0.0650 ซึ่งมีค่าน้อยมาก แสดงว่าค่าที่ทำนายได้ใกล้เคียงกับค่าที่แท้จริง และมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) คือ $(0.2550)^2 = 0.0650$ ซึ่งมีค่าน้อยมาก แสดงว่าตัวแบบมีความถูกต้องดี

ตารางที่ 4.44 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการทำนายการเป็นโรคเบาหวาน ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	1	1	0.935	1	0.966	Negative
	TN Rate	FN Rate	Precision	Recall	F-Measure	Class
	0	0	0	0	0	Positive
Weighted Avg.	0.935	0.935	0.874	0.935	0.904	

จากตารางที่ 4.44 สำหรับคำตอบไม่เป็นโรคเบาหวาน (Class = Negative) มีค่าอัตราความถูกต้องเชิงบวก = 1 ค่าอัตราความผิดพลาดเชิงบวก = 1 ค่าความแม่นยำ = 0.935 ค่าความระลึก = 1 และค่าความถ่วงดุล = 0.966 ส่วนคำตอบเป็นโรคเบาหวาน (Class = Positive) มีค่าอัตราความถูกต้องเชิงลบ = 0 ค่าอัตราความผิดพลาดเชิงลบ = 0 ค่าความแม่นยำ = 0 ค่าความระลึก = 0 และค่าความถ่วงดุล = 0

ตารางที่ 4.45 ผลในส่วนของเมตริกซ์ความสับสนจากการทำนายการเป็นโรคเบาหวาน

ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน

		ผลการจำแนกเป็นโรคเบาหวาน	
		ไม่เป็น	เป็น
ค่าที่แท้จริง	ไม่เป็น	115	0
	เป็น	8	0

จากตารางที่ 4.45 มีข้อมูล 123 คน ตัวแบบสามารถทำนายข้อมูลได้ถูกต้อง 115 คน โดยมีจำนวนข้อมูลที่จำแนกถูกว่าไม่เป็นโรคเบาหวาน 115 คน และจำนวนข้อมูลที่จำแนกถูกว่าเป็นโรคเบาหวาน 0 คน ตัวแบบทำนายข้อมูลไม่ถูกต้อง 8 คน โดยมีจำนวนข้อมูลที่จำแนกผิดว่าเป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วไม่เป็นโรคเบาหวาน 0 คน และจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วเป็นโรคเบาหวาน 8 คน

4.1.5 วิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม (Binary Logistic Regression Method)

เนื่องจากข้อมูลการเป็นโรคเบาหวานมีตัวแปรอิสระหลายตัวและมีตัวแปรตามเพียงตัวเดียวเป็นตัวแปรเชิงคุณภาพที่มีค่าได้เพียง 2 ค่า คือ เป็นและไม่เป็นโรคเบาหวาน จึงเลือกใช้วิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม ได้ผลการทำนายการเป็นโรคเบาหวานดังนี้

4.1.5.1 การสร้างตัวแบบการเป็นโรคเบาหวาน

สร้างตัวแบบการเป็นโรคเบาหวานจากชุดข้อมูลฝึกหัด ด้วยไฟล์ Diabetes-70.csv จำนวน 863 คน ได้ผลดังนี้

ตารางที่ 4.46 ผลในส่วนของการสรุปผลจากการสร้างตัวแบบการเป็นโรคเบาหวาน

ด้วยวิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม

Correctly Classified Instances	789	91.4253 %
Incorrectly Classified Instances	74	8.5747 %
Kappa statistic	0.2387	
Mean absolute error	0.1322	
Root mean squared error	0.2674	
Relative absolute error	83.8324 %	
Root relative squared error	95.5014 %	
Coverage of cases (0.95 level)	97.6825 %	
Mean rel. region size (0.95 level)	69.9305 %	
Total Number of Instances	863	

จากตารางที่ 4.46 จากข้อมูลจำนวน 863 คน ทำนายข้อมูลถูกต้องจำนวน 789 คน คิดเป็น 91.4253% และทำนายข้อมูลไม่ถูกต้องจำนวน 74 คน คิดเป็น 8.5747% มีค่าสถิติแคปปา (Kappa statistic) คือ 0.2387 แสดงว่าข้อมูลมีความสอดคล้องค่อนข้างน้อย มีค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) คือ 0.1322 ซึ่งมีค่าค่อนข้างน้อย แสดงว่าค่าที่ทำนายได้ใกล้เคียงกับค่าที่แท้จริงพอสมควร และมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) คือ $(0.2674)^2 = 0.0715$ ซึ่งมีค่าน้อยมาก แสดงว่าตัวแบบมีความถูกต้องดี

ตารางที่ 4.47 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการสร้างตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.982	0.811	0.599	0.928	0.982	Negative
	TN Rate	FN Rate	Precision	Recall	F-Measure	Class
	0.189	0.018	0.500	0.189	0.275	Positive
Weighted Avg.	0.914	0.743	0.891	0.914	0.896	

จากตารางที่ 4.47 สำหรับคำตอบไม่เป็นโรคเบาหวาน (Class = Negative) มีค่าอัตราความถูกต้องเชิงบวก = 0.982 ค่าอัตราความผิดพลาดเชิงบวก = 0.811 ค่าความแม่นยำ = 0.599 ค่าความระลึก = 0.928 และค่าความถ่วงดุล = 0.982 ส่วนคำตอบเป็นโรคเบาหวาน (Class = Positive) มีค่าอัตราความถูกต้องเชิงลบ = 0.189 ค่าอัตราความผิดพลาดเชิงลบ = 0.018 ค่าความแม่นยำ = 0.500 ค่าความระลึก = 0.189 และค่าความถ่วงดุล = 0.275

ตารางที่ 4.48 ผลในส่วนของเมตริกซ์ความสับสน จากการสร้างตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม

		ผลการจำแนกเป็นโรคเบาหวาน	
		ไม่เป็น	เป็น
ค่าที่แท้จริง	ไม่เป็น	775	14
	เป็น	60	14

จากตารางที่ 4.48 มีข้อมูล 863 คน ตัวแบบสามารถทำนายข้อมูลได้ถูกต้อง 789 คน โดยมีจำนวนข้อมูลที่จำแนกถูกว่าไม่เป็นโรคเบาหวาน 775 คน และจำนวนข้อมูลที่จำแนกถูกว่าเป็นโรคเบาหวาน 14 คน ตัวแบบทำนายข้อมูลไม่ถูกต้อง 74 คน โดยมีจำนวนข้อมูลที่จำแนกผิดว่าเป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วไม่เป็นโรคเบาหวาน 14 คน และจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วเป็นโรคเบาหวาน 60 คน

4.1.5.2 การทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน

ทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวานจากชุดข้อมูลทดสอบประสิทธิภาพ ด้วยไฟล์ Diabetes-20.csv จำนวน 247 คน ได้ผลดังนี้

ตารางที่ 4.49 ผลในส่วนของการสรุปผลจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม

Correctly Classified Instances	226	91.498 %
Incorrectly Classified Instances	21	8.502 %
Kappa statistic	0.0529	
Mean absolute error	0.1319	
Root mean squared error	0.2692	
Coverage of cases (0.95 level)	97.9757 %	
Total Number of Instances	247	

จากตารางที่ 4.49 จากข้อมูลจำนวน 247 คน ทำนายข้อมูลถูกต้องจำนวน 226 คน คิดเป็น 91.498% และทำนายข้อมูลไม่ถูกต้องจำนวน 21 คน คิดเป็น 8.502% มีค่าสถิติแคปปา (Kappa statistic) คือ 0.0529 แสดงว่าข้อมูลมีความสอดคล้องกันน้อยมาก มีค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) คือ 0.1319 ซึ่งมีค่าค่อนข้างน้อย แสดงว่าค่าที่ทำนายได้ใกล้เคียงกับค่าที่แท้จริงพอสมควร และมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) คือ $(0.2692)^2 = 0.0725$ ซึ่งมีค่าน้อยมาก แสดงว่าตัวแบบมีความถูกต้องดี

ตารางที่ 4.50 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.978	0.941	0.934	0.978	0.955	Negative
	TN Rate	FN Rate	Precision	Recall	F-Measure	Class
	0.059	0.022	0.167	0.059	0.087	Positive
Weighted Avg.	0.915	0.878	0.881	0.915	0.896	

จากตารางที่ 4.50 สำหรับคำตอบไม่เป็นโรคเบาหวาน (Class = Negative) มีค่าอัตราความถูกต้องเชิงบวก = 0.978 ค่าอัตราความผิดพลาดเชิงบวก = 0.941 ค่าความแม่นยำ = 0.934 ค่าความระลึก = 0.978 และค่าความถ่วงดุล = 0.955 ส่วนคำตอบเป็นโรคเบาหวาน (Class = Positive) มีค่าอัตราความถูกต้องเชิงลบ = 0.059 ค่าอัตราความผิดพลาดเชิงลบ = 0.022 ค่าความแม่นยำ = 0.167 ค่าความระลึก = 0.059 และค่าความถ่วงดุล = 0.087

ตารางที่ 4.51 ผลในส่วนของเมตริกซ์ความสับสนจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม

		ผลการจำแนกเป็นโรคเบาหวาน	
		ไม่เป็น	เป็น
ค่าที่แท้จริง	ไม่เป็น	225	5
	เป็น	16	1

จากตารางที่ 4.51 มีข้อมูล 247 คน ตัวแบบสามารถทำนายข้อมูลได้ถูกต้อง 226 คน โดยมีจำนวนข้อมูลที่จำแนกถูกว่าไม่เป็นโรคเบาหวาน 225 คน และจำนวนข้อมูลที่จำแนกถูกว่าเป็นโรคเบาหวาน 1 คน ตัวแบบทำนายข้อมูลไม่ถูกต้อง 21 คน โดยมีจำนวนข้อมูลที่จำแนกผิดว่าเป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วไม่เป็นโรคเบาหวาน 5 คน และจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วเป็นโรคเบาหวาน 16 คน

4.1.5.3 การทำนายการเป็นโรคเบาหวาน

การทำนายการเป็นโรคเบาหวานจากชุดข้อมูลทดสอบ ด้วยไฟล์ Diabetes-10.csv จำนวน 123 คน ได้ผลดังนี้

ตารางที่ 4.52 ผลในส่วนของการทำนายชุดข้อมูลทดสอบ ด้วยวิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม

ตัวอย่าง (instance)	ค่าที่แท้จริง (actual)	ค่าที่ทำนายได้ (predicted)	ความคลาดเคลื่อน (error)	การทำนาย (prediction)
1	1:Negative	1:Negative		0.931
2	1:Negative	1:Negative		0.936
3	2:Positive	1:Negative	+	0.898
4	1:Negative	1:Negative		0.951
5	1:Negative	1:Negative		0.993
6	2:Positive	2:Positive		0.524
7	1:Negative	1:Negative		0.988
8	1:Negative	1:Negative		0.979
9	1:Negative	1:Negative		0.991
10	1:Negative	1:Negative		0.784
11	1:Negative	1:Negative		0.969
12	1:Negative	1:Negative		0.952
13	1:Negative	1:Negative		0.991
14	1:Negative	1:Negative		0.999
15	1:Negative	1:Negative		0.937
16	1:Negative	1:Negative		0.955
17	1:Negative	1:Negative		0.509
18	1:Negative	1:Negative		0.977
19	1:Negative	1:Negative		0.983
20	1:Negative	1:Negative		0.97
21	1:Negative	1:Negative		0.99
22	1:Negative	1:Negative		0.987
23	1:Negative	1:Negative		0.743
24	1:Negative	1:Negative		0.947
25	1:Negative	1:Negative		0.95
26	1:Negative	1:Negative		0.979
27	1:Negative	1:Negative		0.989
28	1:Negative	1:Negative		0.875
29	1:Negative	1:Negative		0.973
30	1:Negative	1:Negative		0.965
31	1:Negative	1:Negative		0.98
32	1:Negative	1:Negative		0.947
33	1:Negative	1:Negative		0.977
34	1:Negative	1:Negative		0.944
35	1:Negative	1:Negative		0.954
36	1:Negative	1:Negative		0.984
37	1:Negative	1:Negative		0.947
38	1:Negative	1:Negative		0.948
39	1:Negative	1:Negative		0.996
40	1:Negative	1:Negative		0.984
41	1:Negative	1:Negative		0.986
42	1:Negative	1:Negative		0.96
43	1:Negative	1:Negative		0.984
44	1:Negative	1:Negative		0.983
45	1:Negative	1:Negative		0.923
46	1:Negative	1:Negative		0.961

ตารางที่ 4.52 ผลในส่วนของการทำนายชุดข้อมูลทดสอบ ด้วยวิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม (ต่อ)

ตัวอย่าง (instance)	ค่าที่แท้จริง (actual)	ค่าที่ทำนายได้ (predicted)	ความคลาดเคลื่อน (error)	การทำนาย (prediction)
47	2:Positive	1:Negative	+	0.635
48	1:Negative	1:Negative		0.977
49	1:Negative	1:Negative		0.987
50	1:Negative	1:Negative		0.915
51	1:Negative	1:Negative		0.97
52	1:Negative	1:Negative		0.918
53	1:Negative	1:Negative		0.874
54	1:Negative	1:Negative		0.989
55	1:Negative	1:Negative		0.936
56	1:Negative	1:Negative		0.996
57	1:Negative	1:Negative		0.995
58	1:Negative	1:Negative		0.992
59	1:Negative	2:Positive	+	0.608
60	1:Negative	1:Negative		0.99
61	2:Positive	1:Negative	+	0.76
62	1:Negative	1:Negative		0.994
63	1:Negative	1:Negative		0.966
64	1:Negative	1:Negative		0.991
65	2:Positive	1:Negative	+	0.621
66	1:Negative	1:Negative		0.863
67	1:Negative	1:Negative		0.919
68	1:Negative	1:Negative		0.949
69	2:Positive	1:Negative	+	0.944
70	1:Negative	1:Negative		0.719
71	1:Negative	1:Negative		0.908
72	1:Negative	1:Negative		0.959
73	1:Negative	1:Negative		0.998

ตัวอย่าง (instance)	ค่าที่แท้จริง (actual)	ค่าที่ทำนายได้ (predicted)	ความคลาดเคลื่อน (error)	การทำนาย (prediction)
74	2:Positive	1:Negative	+	0.653
75	1:Negative	1:Negative		0.989
76	1:Negative	1:Negative		0.986
77	1:Negative	1:Negative		0.964
78	1:Negative	1:Negative		0.96
79	1:Negative	1:Negative		0.962
80	1:Negative	1:Negative		0.989
81	1:Negative	1:Negative		0.993
82	2:Positive	2:Positive		0.754
83	1:Negative	1:Negative		0.989
84	1:Negative	1:Negative		0.986
85	1:Negative	1:Negative		0.984
86	1:Negative	1:Negative		0.992
87	1:Negative	1:Negative		0.972
88	1:Negative	1:Negative		0.986
89	1:Negative	1:Negative		0.995
90	1:Negative	1:Negative		0.976
91	1:Negative	1:Negative		0.996
92	1:Negative	1:Negative		0.991
93	1:Negative	1:Negative		0.949
94	1:Negative	1:Negative		0.987
95	1:Negative	1:Negative		0.802
96	1:Negative	1:Negative		0.942
97	1:Negative	1:Negative		0.993
98	1:Negative	1:Negative		0.99
99	1:Negative	1:Negative		0.966
100	1:Negative	1:Negative		0.995

ตารางที่ 4.52 ผลในส่วนของการทำนายชุดข้อมูลทดสอบ ด้วยวิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม (ต่อ)

ตัวอย่าง (instance)	ค่าที่แท้จริง (actual)	ค่าที่ทำนายได้ (predicted)	ความคลาดเคลื่อน (error)	การทำนาย (prediction)
101	1:Negative	1:Negative		0.989
102	1:Negative	1:Negative		0.798
103	1:Negative	1:Negative		0.973
104	1:Negative	1:Negative		0.982
105	1:Negative	1:Negative		0.998
106	1:Negative	1:Negative		0.972
107	1:Negative	1:Negative		0.93
108	1:Negative	1:Negative		0.996
109	1:Negative	1:Negative		0.981
110	1:Negative	1:Negative		0.915
111	1:Negative	1:Negative		0.955
112	1:Negative	1:Negative		0.889

ตัวอย่าง (instance)	ค่าที่แท้จริง (actual)	ค่าที่ทำนายได้ (predicted)	ความคลาดเคลื่อน (error)	การทำนาย (prediction)
113	1:Negative	1:Negative		0.777
114	1:Negative	1:Negative		0.992
115	1:Negative	1:Negative		0.951
116	1:Negative	1:Negative		0.981
117	1:Negative	1:Negative		0.921
118	1:Negative	1:Negative		0.978
119	1:Negative	1:Negative		0.986
120	1:Negative	1:Negative		0.976
121	1:Negative	1:Negative		0.991
122	1:Negative	1:Negative		0.952
123	1:Negative	1:Negative		0.978

จากตารางที่ 4.52 ค่าทำนายที่ทายผิดในช่อง error จะมีเครื่องหมาย + อยู่จำนวน 7 ค่า คือค่าที่ 3, 47, 59, 61, 65, 69 และ 74 โดยที่ค่าที่ 3, 47, 61, 65, 69 และ 74 เป็นโรคเบาหวาน (class 2 : Positive) แต่ทำนายว่าไม่เป็นโรคเบาหวาน (class 1 : Negative) และค่าที่ 59 ไม่เป็นโรคเบาหวาน แต่ทำนายว่าเป็นโรคเบาหวาน

ตารางที่ 4.53 ผลในส่วนของการสรุปผลจากการทำนายการเป็นโรคเบาหวาน ด้วยวิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม

Correctly Classified Instances	116	94.3089 %
Incorrectly Classified Instances	7	5.6911 %
Kappa statistic	0.3402	
Mean absolute error	0.0910	
Root mean squared error	0.1999	
Coverage of cases (0.95 level)	100 %	
Total Number of Instances	123	

จากตารางที่ 4.53 โดยจากข้อมูลจำนวน 123 คน ทำนายข้อมูลถูกต้องจำนวน 116 คน คิดเป็น 94.3089% และทำนายข้อมูลไม่ถูกต้องจำนวน 7 คน คิดเป็น 5.6911% มีค่าสถิติแคปปา (Kappa statistic) คือ 0.3402 แสดงว่าข้อมูลมีความสอดคล้องกันพอสมควร มีค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) คือ 0.0910 ซึ่งมีค่าน้อยมาก แสดงว่าค่าที่ทำนายได้ใกล้เคียงกับค่าที่แท้จริง และมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) คือ $(0.1999)^2 = 0.0400$ ซึ่งมีค่าน้อยมาก แสดงว่าตัวแบบมีความถูกต้องดี

ตารางที่ 4.54 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการทำนายการเป็นโรคเบาหวาน ด้วยวิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.991	0.750	0.950	0.991	0.970	Negative
	TN Rate	FN Rate	Precision	Recall	F-Measure	Class
	0.250	0.009	0.667	0.250	0.364	Positive
Weighted Avg.	0.943	0.702	0.932	0.943	0.931	

จากตารางที่ 4.54 สำหรับคำตอบไม่เป็นโรคเบาหวาน (Class = Negative) มีค่าอัตราความถูกต้องเชิงบวก = 0.991 ค่าอัตราความผิดพลาดเชิงบวก = 0.750 ค่าความแม่นยำ = 0.950 ค่าความระลึก = 0.991 และค่าความถ่วงดุล = 0.970 ส่วนคำตอบเป็นโรคเบาหวาน (Class = Positive) มีค่าอัตราความถูกต้องเชิงลบ = 0.250 ค่าอัตราความผิดพลาดเชิงลบ = 0.009 ค่าความแม่นยำ = 0.667 ค่าความระลึก = 0.250 และค่าความถ่วงดุล = 0.364

ตารางที่ 4.55 ผลในส่วนของเมตริกซ์ความสับสนจากการทำนายการเป็นโรคเบาหวาน

ด้วยวิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม

		ผลการจำแนกเป็นโรคเบาหวาน	
		ไม่เป็น	เป็น
ค่าที่แท้จริง	ไม่เป็น	114	1
	เป็น	6	2

จากตารางที่ 4.55 มีข้อมูล 123 คน ตัวแบบสามารถทำนายข้อมูลได้ถูกต้อง 116 คน โดยมีจำนวนข้อมูลที่จำแนกถูกว่าไม่เป็นโรคเบาหวาน 114 คน และจำนวนข้อมูลที่จำแนกถูกว่าเป็นโรคเบาหวาน 2 คน ตัวแบบทำนายข้อมูลไม่ถูกต้อง 7 คน โดยมีจำนวนข้อมูลที่จำแนกผิดว่าเป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วไม่เป็นโรคเบาหวาน 1 คน และจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วเป็นโรคเบาหวาน 6 คน

4.1.6 วิธีนาอีฟ เบย์ (Naïve Bayes Method)

เนื่องจากข้อมูลการเป็นโรคเบาหวานมีตัวแปรอิสระหลายตัวและมีตัวแปรตามเพียงตัวเดียวเป็นตัวแปรเชิงคุณภาพที่มีค่าได้เพียง 2 ค่า คือ เป็นและไม่เป็นโรคเบาหวาน จึงเลือกใช้วิธีนาอีฟ เบย์ ได้ผลการทำนายการเป็นโรคเบาหวานดังนี้

4.1.6.1 การสร้างตัวแบบการเป็นโรคเบาหวาน

สร้างตัวแบบการเป็นโรคเบาหวานจากชุดข้อมูลฝึกหัด ด้วยไฟล์ Diabetes-70.csv จำนวน 863 คน ได้ผลดังนี้

ตารางที่ 4.56 ผลในส่วนของการสรุปผลจากการสร้างตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีนาอีฟ เบย์

Correctly Classified Instances	762	88.2966 %
Incorrectly Classified Instances	101	11.7034 %
Kappa statistic	0.3164	
Mean absolute error	0.1379	
Root mean squared error	0.3062	
Relative absolute error	87.5219 %	
Root relative squared error	109.3606 %	
Coverage of cases (0.95 level)	95.4809 %	
Mean rel. region size (0.95 level)	62.3407 %	
Total Number of Instances	863	

จากตารางที่ 4.56 จากข้อมูลจำนวน 863 คน ทำนายข้อมูลถูกต้องจำนวน 762 คน คิดเป็น 88.2966% และทำนายข้อมูลไม่ถูกต้องจำนวน 101 คน คิดเป็น 11.7034% มีค่าสถิติแคปปา (Kappa statistic) คือ 0.3164 แสดงว่าข้อมูลมีความสอดคล้องพอสมควร มีค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) คือ 0.1379 ซึ่งมีค่าค่อนข้างน้อย แสดงว่าค่าที่ทำนายได้ใกล้เคียงกับค่าที่แท้จริงพอสมควร และมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) คือ $(0.3062)^2 = 0.0938$ ซึ่งมีค่าน้อยมาก แสดงว่าตัวแบบมีความถูกต้องดี

ตารางที่ 4.57 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการสร้างตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีนาอีฟ เบย์

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.926	0.581	0.944	0.926	0.935	Negative
	TN Rate	FN Rate	Precision	Recall	F-Measure	Class
	0.419	0.074	0.348	0.419	0.380	Positive
Weighted Avg.	0.883	0.538	0.893	0.883	0.888	

จากตารางที่ 4.57 สำหรับคำตอบไม่เป็นโรคเบาหวาน (Class = Negative) มีค่าอัตราความถูกต้องเชิงบวก = 0.926 ค่าอัตราความผิดพลาดเชิงบวก = 0.581 ค่าความแม่นยำ = 0.944 ค่าความระลึก = 0.926 และค่าความถ่วงดุล = 0.935 ส่วนคำตอบเป็นโรคเบาหวาน (Class = Positive) มีค่าอัตราความถูกต้องเชิงลบ = 0.419 ค่าอัตราความผิดพลาดเชิงลบ = 0.074 ค่าความแม่นยำ = 0.348 ค่าความระลึก = 0.419 และค่าความถ่วงดุล = 0.380

ตารางที่ 4.58 ผลในส่วนของเมตริกซ์ความสับสน จากการสร้างตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีนาอีฟ เบย์

		ผลการจำแนกเป็นโรคเบาหวาน	
		ไม่เป็น	เป็น
ค่าที่แท้จริง	ไม่เป็น	731	58
	เป็น	43	31

จากตารางที่ 4.58 มีข้อมูล 863 คน ตัวแบบสามารถทำนายข้อมูลได้ถูกต้อง 762 คน โดยมีจำนวนข้อมูลที่จำแนกถูกว่าไม่เป็นโรคเบาหวาน 731 คน และจำนวนข้อมูลที่จำแนกถูกว่าเป็นโรคเบาหวาน 31 คน ตัวแบบทำนายข้อมูลไม่ถูกต้อง 101 คน โดยมีจำนวนข้อมูลที่จำแนกผิดว่าเป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วไม่เป็นโรคเบาหวาน 58 คน และจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วเป็นโรคเบาหวาน 43 คน

4.1.6.2 การทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน

ทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวานจากชุดข้อมูลทดสอบประสิทธิภาพ ด้วยไฟล์ Diabetes-20.csv จำนวน 247 คน ได้ผลดังนี้

ตารางที่ 4.59 ผลในส่วนของการสรุปผลจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีนาอีฟ เบย์

Correctly Classified Instances	212	85.83 %
Incorrectly Classified Instances	35	14.17 %
Kappa statistic	0.1121	
Mean absolute error	0.1634	
Root mean squared error	0.3375	
Coverage of cases (0.95 level)	93.9271 %	
Total Number of Instances	247	

จากตารางที่ 4.59 จากข้อมูลจำนวน 247 คน ทำนายข้อมูลถูกต้องจำนวน 212 คน คิดเป็น 85.83% และทำนายข้อมูลไม่ถูกต้องจำนวน 35 คน คิดเป็น 14.17% มีค่าสถิติแคปปา (Kappa statistic) คือ 0.1129 แสดงว่าข้อมูลมีความสอดคล้องกันค่อนข้างน้อย มีค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) คือ 0.1634 ซึ่งมีค่าค่อนข้างน้อย แสดงว่าค่าที่ทำนายได้ใกล้เคียงกับค่าที่แท้จริงพอสมควร และมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) คือ $(0.3375)^2 = 0.1139$ ซึ่งมีค่าค่อนข้างน้อย แสดงว่าตัวแบบมีความถูกต้องพอสมควร

ตารางที่ 4.60 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีนาอีฟ เบย์

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.904	0.765	0.941	0.904	0.922	Negative
	TN Rate	FN Rate	Precision	Recall	F-Measure	Class
	0.235	0.096	0.154	0.235	0.186	Positive
Weighted Avg.	0.858	0.719	0.887	0.858	0.872	

จากตารางที่ 4.60 สำหรับคำตอบไม่เป็นโรคเบาหวาน (Class = Negative) มีค่าอัตราความถูกต้องเชิงบวก = 0.904 ค่าอัตราความผิดพลาดเชิงบวก = 0.765 ค่าความแม่นยำ = 0.941 ค่าความระลึก = 0.904 และค่าความถ่วงดุล = 0.922 ส่วนคำตอบเป็นโรคเบาหวาน (Class = Positive) มีค่าอัตราความถูกต้องเชิงลบ = 0.235 ค่าอัตราความผิดพลาดเชิงลบ = 0.096 ค่าความแม่นยำ = 0.154 ค่าความระลึก = 0.235 และค่าความถ่วงดุล = 0.186

ตารางที่ 4.61 ผลในส่วนของเมตริกซ์ความสับสนจากการทดสอบความถูกต้องของตัวแบบการเป็นโรคเบาหวาน ด้วยวิธีนาอีฟ เบย์

		ผลการจำแนกเป็นโรคเบาหวาน	
		ไม่เป็น	เป็น
ค่าที่แท้จริง	ไม่เป็น	208	22
	เป็น	13	4

จากตารางที่ 4.61 มีข้อมูล 247 คน ตัวแบบสามารถทำนายข้อมูลได้ถูกต้อง 212 คน โดยมีจำนวนข้อมูลที่จำแนกถูกว่าไม่เป็นโรคเบาหวาน 208 คน และจำนวนข้อมูลที่จำแนกถูกว่าเป็นโรคเบาหวาน 4 คน ตัวแบบทำนายข้อมูลไม่ถูกต้อง 35 คน โดยมีจำนวนข้อมูลที่จำแนกผิดว่าเป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วไม่เป็นโรคเบาหวาน 22 คน และจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วเป็นโรคเบาหวาน 13 คน

4.1.6.3 การทำนายการเป็นโรคเบาหวาน

การทำนายการเป็นโรคเบาหวานจากชุดข้อมูลทดสอบ ด้วยไฟล์ Diabetes-10.csv จำนวน 123 คน ได้ผลดังนี้

ตารางที่ 4.62 ผลในส่วนของการทำนายชุดข้อมูลทดสอบ ด้วยวิธีนาอีฟ เบย์

ตัวอย่าง (instance)	ค่าที่แท้จริง (actual)	ค่าที่ทำนายได้ (predicted)	ความคลาดเคลื่อน (error)	การทำนาย (prediction)
1	1:Negative	1:Negative		0.987
2	1:Negative	1:Negative		0.913
3	2:Positive	1:Negative	+	0.87
4	1:Negative	1:Negative		0.572
5	1:Negative	1:Negative		0.999
6	2:Positive	2:Positive		0.944
7	1:Negative	1:Negative		0.997
8	1:Negative	1:Negative		0.993
9	1:Negative	1:Negative		0.999
10	1:Negative	2:Positive	+	0.948
11	1:Negative	1:Negative		0.789
12	1:Negative	1:Negative		0.698
13	1:Negative	1:Negative		0.998
14	1:Negative	1:Negative		0.999
15	1:Negative	1:Negative		0.892
16	1:Negative	1:Negative		0.995
17	1:Negative	1:Negative		0.969
18	1:Negative	1:Negative		0.986
19	1:Negative	1:Negative		0.994
20	1:Negative	1:Negative		0.993
21	1:Negative	1:Negative		0.993
22	1:Negative	1:Negative		0.998
23	1:Negative	1:Negative		0.945

ตัวอย่าง (instance)	ค่าที่แท้จริง (actual)	ค่าที่ทำนายได้ (predicted)	ความคลาดเคลื่อน (error)	การทำนาย (prediction)
24	1:Negative	1:Negative		0.972
25	1:Negative	1:Negative		0.996
26	1:Negative	1:Negative		0.926
27	1:Negative	1:Negative		0.997
28	1:Negative	1:Negative		0.69
29	1:Negative	1:Negative		0.995
30	1:Negative	1:Negative		0.961
31	1:Negative	1:Negative		0.668
32	1:Negative	1:Negative		0.994
33	1:Negative	1:Negative		0.851
34	1:Negative	1:Negative		0.875
35	1:Negative	1:Negative		0.997
36	1:Negative	1:Negative		0.993
37	1:Negative	1:Negative		0.562
38	1:Negative	1:Negative		0.992
39	1:Negative	1:Negative		0.999
40	1:Negative	1:Negative		0.998
41	1:Negative	1:Negative		0.963
42	1:Negative	1:Negative		0.959
43	1:Negative	1:Negative		0.993
44	1:Negative	1:Negative		0.962
45	1:Negative	1:Negative		0.99
46	1:Negative	1:Negative		0.993

ตารางที่ 4.62 ผลในส่วนของการทำนายชุดข้อมูลทดสอบ ด้วยวิธีนาอีฟ เบย์ (ต่อ)

ตัวอย่าง (instance)	ค่าที่แท้จริง (actual)	ค่าที่ทำนายได้ (predicted)	ความคลาดเคลื่อน (error)	การทำนาย (prediction)
47	2:Positive	2:Positive		0.622
48	1:Negative	1:Negative		0.997
49	1:Negative	1:Negative		0.998
50	1:Negative	1:Negative		0.812
51	1:Negative	1:Negative		0.998
52	1:Negative	1:Negative		0.996
53	1:Negative	1:Negative		0.928
54	1:Negative	1:Negative		0.998
55	1:Negative	1:Negative		0.965
56	1:Negative	1:Negative		0.999
57	1:Negative	1:Negative		0.999
58	1:Negative	1:Negative		0.999
59	1:Negative	2:Positive	+	0.989
60	1:Negative	1:Negative		0.992
61	2:Positive	2:Positive		0.716
62	1:Negative	1:Negative		0.999
63	1:Negative	1:Negative		0.989
64	1:Negative	1:Negative		0.992
65	2:Positive	2:Positive		0.635
66	1:Negative	1:Negative		0.853
67	1:Negative	2:Positive	+	0.996
68	1:Negative	1:Negative		0.995
69	2:Positive	1:Negative	+	0.936
70	1:Negative	1:Negative		0.685
71	1:Negative	1:Negative		0.618
72	1:Negative	2:Positive	+	0.914
73	1:Negative	1:Negative		0.999

ตัวอย่าง (instance)	ค่าที่แท้จริง (actual)	ค่าที่ทำนายได้ (predicted)	ความคลาดเคลื่อน (error)	การทำนาย (prediction)
74	2:Positive	1:Negative	+	0.786
75	1:Negative	1:Negative		0.996
76	1:Negative	1:Negative		0.998
77	1:Negative	1:Negative		0.532
78	1:Negative	1:Negative		0.995
79	1:Negative	1:Negative		0.997
80	1:Negative	1:Negative		0.992
81	1:Negative	1:Negative		0.999
82	2:Positive	2:Positive		0.988
83	1:Negative	1:Negative		0.998
84	1:Negative	1:Negative		0.988
85	1:Negative	1:Negative		0.997
86	1:Negative	1:Negative		0.989
87	1:Negative	1:Negative		0.969
88	1:Negative	1:Negative		0.996
89	1:Negative	1:Negative		0.998
90	1:Negative	1:Negative		0.993
91	1:Negative	1:Negative		0.999
92	1:Negative	1:Negative		0.999
93	1:Negative	1:Negative		0.813
94	1:Negative	1:Negative		0.998
95	1:Negative	1:Negative		0.915
96	1:Negative	2:Positive	+	0.996
97	1:Negative	1:Negative		0.989
98	1:Negative	1:Negative		0.999
99	1:Negative	1:Negative		0.998
100	1:Negative	1:Negative		0.999

ตารางที่ 4.62 ผลในส่วนของการทำนายชุดข้อมูลทดสอบ ด้วยวิธีนาอีฟ เบย์ (ต่อ)

ตัวอย่าง (instance)	ค่าที่แท้จริง (actual)	ค่าที่ทำนายได้ (predicted)	ความคลาดเคลื่อน (error)	การทำนาย (prediction)
101	1:Negative	1:Negative		0.999
102	1:Negative	1:Negative		0.953
103	1:Negative	1:Negative		0.997
104	1:Negative	1:Negative		0.999
105	1:Negative	1:Negative		0.999
106	1:Negative	1:Negative		0.997
107	1:Negative	1:Negative		0.944
108	1:Negative	1:Negative		0.998
109	1:Negative	1:Negative		0.934
110	1:Negative	1:Negative		0.996
111	1:Negative	1:Negative		0.974
112	1:Negative	2:Positive	+	0.88

ตัวอย่าง (instance)	ค่าที่แท้จริง (actual)	ค่าที่ทำนายได้ (predicted)	ความคลาดเคลื่อน	การทำนาย (prediction)
113	1:Negative	2:Positive	+	1
114	1:Negative	1:Negative		0.997
115	1:Negative	1:Negative		0.991
116	1:Negative	1:Negative		0.997
117	1:Negative	1:Negative		0.961
118	1:Negative	1:Negative		0.996
119	1:Negative	1:Negative		0.995
120	1:Negative	1:Negative		0.992
121	1:Negative	1:Negative		0.998
122	1:Negative	1:Negative		0.978
123	1:Negative	2:Positive	+	0.579

จากตารางที่ 4.62 ค่าทำนายที่ทายผิดในช่อง error จะมีเครื่องหมาย + อยู่จำนวน 11 ค่า คือค่าที่ 3, 10, 59, 67, 69, 72, 74, 96, 112, 113 และ 123 โดยที่ค่าที่ 3, 69 และ 74 เป็นโรคเบาหวาน (class 2 : Positive) แต่ทำนายว่าไม่เป็นโรคเบาหวาน (class 1 : Negative) และค่าที่ 10, 59, 67, 72, 96, 112, 113 และ 123 ไม่เป็นโรคเบาหวาน แต่ทำนายว่าเป็นโรคเบาหวาน

ตารางที่ 4.63 ผลในส่วนของการสรุปผลจากการทำนายการเป็นโรคเบาหวาน ด้วยวิธีนาอีฟ เบย์

Correctly Classified Instances	112	91.0569 %
Incorrectly Classified Instances	11	8.9431 %
Kappa statistic	0.4303	
Mean absolute error	0.1325	
Root mean squared error	0.2962	
Coverage of cases (0.95 level)	96.748 %	
Total Number of Instances	123	

จากตารางที่ 4.63 โดยจากข้อมูลจำนวน 123 คน ทำนายข้อมูลถูกต้องจำนวน 112 คน คิดเป็น 91.0569% และทำนายข้อมูลไม่ถูกต้องจำนวน 11 คน คิดเป็น 8.9431% มีค่าสถิติแคปปา (Kappa statistic) คือ 0.4303 แสดงว่าข้อมูลมีความสอดคล้องกันปานกลาง มีค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) คือ 0.1325 ซึ่งมีค่าค่อนข้างน้อย แสดงว่าค่าที่ทำนายได้ใกล้เคียงกับค่าที่แท้จริงพอสมควร และมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) คือ $(0.2962)^2$ = 0.0877 ซึ่งมีค่าน้อยมาก แสดงว่าตัวแบบมีความถูกต้องดี

ตารางที่ 4.64 ผลในส่วนของความถูกต้องของรายละเอียดในแต่ละคำตอบจากการทำนายการเป็นโรคเบาหวาน ด้วยวิธีนาอีฟ เบย์

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.930	0.375	0.973	0.930	0.961	Negative
	TN Rate	FN Rate	Precision	Recall	F-Measure	Class
	0.625	0.070	0.385	0.625	0.476	Positive
Weighted Avg.	0.911	0.355	0.934	0.911	0.920	

จากตารางที่ 4.64 สำหรับคำตอบไม่เป็นโรคเบาหวาน (Class = Negative) มีค่าอัตราความถูกต้องเชิงบวก = 0.930 ค่าอัตราความผิดพลาดเชิงบวก = 0.375 ค่าความแม่นยำ = 0.973 ค่าความระลึก = 0.930 และค่าความถ่วงดุล = 0.961 ส่วนคำตอบเป็นโรคเบาหวาน (Class = Positive) มีค่าอัตราความถูกต้องเชิงลบ = 0.625 ค่าอัตราความผิดพลาดเชิงลบ = 0.070 ค่าความแม่นยำ = 0.385 ค่าความระลึก = 0.625 และค่าความถ่วงดุล = 0.476

ตารางที่ 4.65 ผลในส่วนของเมตริกซ์ความสับสนจากการทำนายการเป็นโรคเบาหวาน ด้วยวิธีนาอีฟ เบย์

		ผลการจำแนกเป็นโรคเบาหวาน	
		ไม่เป็น	เป็น
ค่าที่แท้จริง	ไม่เป็น	107	8
	เป็น	3	5

จากตารางที่ 4.65 มีข้อมูล 123 คน ตัวแบบสามารถทำนายข้อมูลได้ถูกต้อง 112 คน โดยมีจำนวนข้อมูลที่จำแนกถูกว่าไม่เป็นโรคเบาหวาน 107 คน และจำนวนข้อมูลที่จำแนกถูกว่าเป็นโรคเบาหวาน 5 คน ตัวแบบทำนายข้อมูลไม่ถูกต้อง 11 คน โดยมีจำนวนข้อมูลที่จำแนกผิดว่าเป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วไม่เป็นโรคเบาหวาน 8 คน และจำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรคเบาหวาน ซึ่งค่าที่แท้จริงแล้วเป็นโรคเบาหวาน 3 คน

4.2 ผลการเปรียบเทียบประสิทธิภาพในการทำนายผลของวิธีการจำแนกกลุ่ม

จากการเปรียบเทียบประสิทธิภาพในการทำนายผลของวิธีการจำแนกกลุ่มได้ถูกต้องและความแม่นยำในการทำนายการเป็นโรคเบาหวาน ระหว่างวิธีความใกล้เคียงกันมากที่สุด วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ วิธีโครงข่ายประสาท วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม และวิธีนาอีฟ เบย์ ได้ผลดังตารางที่ 4.66 และ 4.67

ตารางที่ 4.66 ผลการเปรียบเทียบประสิทธิภาพในการทำนายผลของวิธีการจำแนกกลุ่มในการเป็นโรคเบาหวานทั้ง 6 วิธี

การเปรียบเทียบประสิทธิภาพในการทำนายผล	ค่าความถูกต้อง (Accuracy)	ค่าอัตราความถูกต้องเชิงบวก (TP Rate)	ค่าอัตราความถูกต้องเชิงลบ (TN Rate)	ค่าความแม่นยำ (Precision)
วิธีความใกล้เคียงกันมากที่สุด	87.80%	0.939	0	0.931
วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ	93.50%	1	0	0.935
วิธีโครงข่ายประสาท	**95.94%**	0.983	**0.625**	**0.974**
วิธีซัพพอร์ตเวกเตอร์แมชชีน	93.50%	1	0	0.935
วิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม	94.31%	0.991	0.250	0.950
วิธีนาอีฟ เบย์	91.06%	0.930	**0.625**	0.973

ตารางที่ 4.66 ผลการเปรียบเทียบประสิทธิภาพในการทำนายผลของวิธีการจำแนกกลุ่มในการเป็นโรคเบาหวานทั้ง 6 วิธี (ต่อ)

การเปรียบเทียบประสิทธิภาพในการทำนายผล	ค่าอัตราความผิดพลาดเชิงบวก (FP Rate)	ค่าอัตราความผิดพลาดเชิงลบ (FN Rate)	ค่าความระลึก (Recall)	ค่าความถ่วงดุล (F-Measure)
วิธีความใกล้เคียงกันมากที่สุด	1	0.061	0.939	0.935
วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ	1	0	1	0.966
วิธีโครงข่ายประสาท	**0.375**	0.017	0.983	**0.978**
วิธีซัพพอร์ตเวกเตอร์แมชชีน	1	0	1	0.966
วิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม	0.750	0.009	0.991	0.970
วิธีนาอีฟ เบย์	**0.375**	0.070	0.930	0.951

ตารางที่ 4.67 ผลการเปรียบเทียบประสิทธิภาพในการทำนายผลของวิธีการจำแนกกลุ่มในการเป็นโรคเบาหวานทั้ง 6 วิธี

การเปรียบเทียบประสิทธิภาพในการทำนายผล	ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE)	ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE)
วิธีความใกล้เคียงกันมากที่สุด	0.1228	0.1212
วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ	0.1137	0.0515
วิธีโครงข่ายประสาท	**0.0491**	**0.0396**
วิธีซัพพอร์ตเวกเตอร์แมชชีน	0.0650	0.0650
วิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม	0.0910	0.0400
วิธีนาอีฟ เบย์	0.1325	0.0877

จากตารางที่ 4.66 พบว่าวิธีโครงข่ายประสาทมีค่าความถูกต้อง ค่าอัตราความถูกต้องเชิงลบ ค่าความแม่นยำ ค่าอัตราความผิดพลาดเชิงบวก และค่าความถ่วงดุลดีที่สุดคือ 95.94%, 0.625, 0.974, 0.375 และ 0.978 ตามลำดับ ส่วนวิธีแผนภาพต้นไม้เพื่อการตัดสินใจและวิธีซัพพอร์ตเวกเตอร์แมชชีนมีค่าความถูกต้องเชิงบวก ค่าอัตราความผิดพลาดเชิงลบและค่าความระลึกดีที่สุดคือ 1, 0 และ 1 ตามลำดับ และวิธีนาอีฟ เบย์มีค่าอัตราความถูกต้องเชิงลบและค่าอัตราความผิดพลาดเชิงบวกที่ดีที่สุด คือ 0.625 และ 0.375 ดังนั้นวิธีโครงข่ายประสาทมีประสิทธิภาพในการทำนายผลดีที่สุด

จากตารางที่ 4.67 พบว่าวิธีโครงข่ายประสาทมีค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ยและความคลาดเคลื่อนกำลังสองเฉลี่ยน้อยที่สุดคือ 0.0491 และ 0.0396 ตามลำดับ แสดงว่ามีความคลาดเคลื่อนน้อยที่สุด ดังนั้นวิธีโครงข่ายประสาทมีประสิทธิภาพในการทำนายผลดีที่สุด

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

ในการทำวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาทำความเข้าใจและเปรียบเทียบประสิทธิภาพในการทำนายผลการเป็นโรคเบาหวาน วิธีการจำแนกกลุ่มที่นำมาเปรียบเทียบประสิทธิภาพในการทำนายผลคือ วิธีความใกล้เคียงกันมากที่สุด โดยใช้อัลกอริทึมชนิด IBk วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ โดยใช้อัลกอริทึมชนิด J48 วิธีโครงข่ายประสาท โดยใช้อัลกอริทึมชนิดเพอร์เซปตรอนแบบหลายชั้น กำหนดค่าอัตราการเรียนรู้เป็น 0.1, 0.2, 0.3, 0.4 และ 0.5 ค่าโมเมนตัมเป็น 0.5, 0.6, 0.7, 0.8 และ 0.9 จำนวนรอบการสอน 20,000 รอบ และชั้นซ่อน 1 ชั้น วิธีซัพพอร์ตเวกเตอร์แมชชีน โดยใช้อัลกอริทึม SMO ชนิดโพลิโนเมียลเคอร์เนล วิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม และวิธีนาอีฟ เบย์ ในการเปรียบเทียบประสิทธิภาพของวิธีการจำแนกกลุ่มทั้ง 6 วิธี จะใช้ค่าความถูกต้อง ค่าอัตราความถูกต้องเชิงบวก ค่าอัตราความถูกต้องเชิงลบ ค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุล โดยจะพิจารณาจากค่าเหล่านี้ที่มีค่ามากที่สุด ส่วนค่าอัตราความผิดพลาดเชิงบวก ค่าอัตราความผิดพลาดเชิงลบ ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) และค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) จะพิจารณาจากค่าเหล่านี้ที่มีค่าน้อยที่สุด พบว่าวิธีโครงข่ายประสาทมีค่าความถูกต้อง ค่าอัตราความถูกต้องเชิงลบ ค่าความแม่นยำ ค่าอัตราความผิดพลาดเชิงบวก ค่าความถ่วงดุล ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย และค่าความคลาดเคลื่อนกำลังสองเฉลี่ยดีที่สุดคือ 95.94%, 0.625, 0.974, 0.375, 0.978, 0.0491 และ 0.0396 ตามลำดับ ส่วนวิธีแผนภาพต้นไม้เพื่อการตัดสินใจและวิธีซัพพอร์ตเวกเตอร์แมชชีนมีค่าอัตราความถูกต้องเชิงบวก ค่าอัตราความผิดพลาดเชิงลบและค่าความระลึกดีที่สุดคือ 1, 0 และ 1 ตามลำดับ และวิธีนาอีฟ เบย์มีค่าอัตราความถูกต้องเชิงลบและค่าอัตราความผิดพลาดเชิงบวกดีที่สุดคือ 0.625 และ 0.375 ตามลำดับ ดังนั้นวิธีโครงข่ายประสาทมีประสิทธิภาพดีที่สุด

5.2 การอภิปรายผล

ผลการศึกษาการเปรียบเทียบประสิทธิภาพในการทำนายผลการเป็นโรคเบาหวาน โดยพิจารณาจากค่าความถูกต้อง ค่าอัตราความถูกต้องเชิงบวก ค่าอัตราความถูกต้องเชิงลบ ค่าความแม่นยำ ค่าอัตราความผิดพลาดเชิงบวก ค่าอัตราความผิดพลาดเชิงลบ ค่าความระลึก และค่าความถ่วงดุลพบว่าวิธีโครงข่ายประสาทมีประสิทธิภาพในการทำนายผลดีที่สุด ซึ่งให้ผลสอดคล้องกับงานวิจัยของกิตติพล วิแสง และคณะ (2552) ที่ได้ศึกษาเรื่องการวิเคราะห์ปัจจัยเสี่ยงของโรคเบาหวาน พบว่าปัจจัยเสี่ยงของโรคเบาหวานมีความสำคัญอย่างยิ่งในการนำมาเป็นเครื่องมือประเมินการเกิดโรคเบาหวาน แทนการตรวจเลือด บทความนี้ศึกษาและวิเคราะห์ปัจจัยเสี่ยงและลำดับของปัจจัยเสี่ยงในการเกิดโรคเบาหวาน โดยตัวแบบที่ใช้ศึกษาคือตัวแบบ Back-propagation Neural Networks, Radial Basis Function Network และตัวแบบ Naïve Bayes ผลการวินิจฉัยทางการแพทย์ได้ถูกนำมาอ้างอิงเพื่อวัดความถูกต้องของตัวแบบ ผลการศึกษาแสดงให้เห็นว่าตัวแบบ Back-propagation Network ให้ผลลัพธ์ที่มีความถูกต้องมากที่สุด ปัจจัยเสี่ยงและลำดับของปัจจัยเสี่ยงสามารถตัดสินได้จากตัวแบบที่ศึกษา ซึ่งข้อมูลเหล่านี้สามารถนำไปสร้างเป็นแบบประเมินความเสี่ยงโรคเบาหวานโดยไม่ต้องอาศัยการตรวจเลือดได้ รวมทั้งเป็นเครื่องมือในการประเมินตนเองและสามารถพัฒนาเป็นระบบผู้เชี่ยวชาญสำหรับการวินิจฉัยโรคเบาหวานต่อไปได้

ยังให้ผลสอดคล้องกับงานวิจัยของชาญชัย พรมโคตร และคณะ (2554) ที่ทำศึกษาการวิเคราะห์ข้อมูลความเสี่ยงของการเป็นโรคหอบหืดด้วยเทคนิคโครงข่ายประสาทเทียมแบบหลายชั้นร่วมกับขั้นตอนวิธีการเรียนรู้แบบแพร่ย้อนกลับ พบว่าโรคหอบหืดเป็นอาการป่วยที่ร้ายแรงอาจทำให้เสียชีวิตถ้าผู้ป่วยไม่ปฏิบัติตัวให้ถูกต้องจนอาการรุนแรง การตรวจผู้ป่วย ทำโดยผู้เชี่ยวชาญประกอบกับผลทางห้องปฏิบัติการในรายที่มีอาการแบบเฉียบพลันจะส่งผลให้การรักษาผู้ป่วยไม่ทันท่วงทีจนเกิดอาการรุนแรงและอาการแทรกซ้อน แต่ถ้าผู้ป่วยทราบว่าตนเองเป็นโรคหอบหืดตั้งแต่ระยะแรกของโรค และได้รับการตรวจรักษา พร้อมทั้งปฏิบัติตนอย่างเหมาะสม จะทำให้ลดความเสียงของการเสียชีวิตและภาวะแทรกซ้อนได้ การจำแนกข้อมูลผู้ป่วยซึ่งเป็น ข้อมูลอาการ เช่น ไอแน่นหน้าอก หายใจหอบเหนื่อย ด้วยเทคนิคโครงข่ายประสาทเทียมแบบหลายชั้น (Multi-Layer Perceptron) ร่วมกับขั้นตอนวิธีการเรียนรู้แบบแพร่ย้อนกลับ (Back Propagation Learning Algorithm) จึงถูกนำเสนอเพื่อวิเคราะห์ความเสียงของการเกิดหอบหืด โดยวิธีการนี้มีความถูกต้องมากกว่าร้อยละ 90 และสามารถนำไปใช้ประกอบการตัดสินใจของผู้ป่วยเพื่อเข้ารับการรักษาตั้งแต่เริ่มต้นได้ นอกจากนี้ยังให้ผลสอดคล้องกับงานวิจัยของสุมนพิชญ์ พลศรี และอัจฉรา มหาวีรวัฒน์ (2554) ที่ศึกษาการวิเคราะห์ความเสี่ยงของการเป็นวัณโรคด้วยเทคนิคโครงข่ายประสาทเทียม พบว่าวัณโรคเป็นโรคติดต่อเรื้อรังรักษาให้หายได้ แต่ถ้าไม่กินยาให้ครบตามกำหนดอาจทำให้เชื้อดื้อยา ยากแก่การรักษา แต่ถ้าหากผู้ป่วยทราบว่าตนเองเป็นวัณโรคในระยะแรก ๆ จะช่วยลดความรุนแรงของโรคก่อนที่จะเป็นอันตรายถึงแก่ชีวิตหรือสามารถป้องกันไม่ให้โรคติดต่อไปสู่บุคคลอื่นได้ วิธีการวิเคราะห์ความเสี่ยงของการเกิด วัณโรคจึงถูกนำเสนอ โดยประยุกต์ใช้เทคนิคโครงข่ายประสาทเทียมแบบหลายชั้น (Multi-Layer Perceptron) ร่วมกับขั้นตอนวิธีการเรียนรู้แบบแพร่ย้อนกลับ (Back Propagation Learning Algorithm) เพื่อจำแนกข้อมูลผู้ป่วย ซึ่งได้แก่ ข้อมูลอาการ ประวัติการเป็นโรคของสมาชิกในครอบครัวและปัจจัยเสี่ยงที่มีผลต่อการเป็นวัณโรค จากผลการวิเคราะห์พบว่าวิธีการนี้มีความถูกต้องแม่นยำและสามารถนำไปใช้ประกอบการตัดสินใจของผู้ป่วยเพื่อเข้ารับการรักษาตั้งแต่ระยะแรกของโรคได้

ส่วนการเปรียบเทียบประสิทธิภาพในการทำนายผลการเป็นโรคเบาหวานโดยพิจารณาจากค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ยและความคลาดเคลื่อนกำลังสองเฉลี่ยพบว่า วิธีโครงข่ายประสาทมีค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ยและความคลาดเคลื่อนกำลังสองเฉลี่ยน้อยที่สุด ดังนั้นวิธีโครงข่ายประสาทมีประสิทธิภาพในการทำนายผลดีที่สุด ซึ่งให้ผลตรงข้ามกับงานวิจัยของณัฐวุฒิ ศิริกุลรุ่งโรจน์ และคณะ (2556) ที่พบว่าวิธีการจำแนกกลุ่มที่มีประสิทธิภาพในการทำนายผลดีที่สุดโดยเปรียบเทียบจากค่าความคลาดเคลื่อนกำลังสองเฉลี่ย คือ วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ ซึ่งอาจจะมีความเป็นไปได้เนื่องจากในงานวิจัยนี้ใช้ข้อมูลเพียงเรื่องเดียวคือ ข้อมูลการเป็นโรคเบาหวาน ถ้าใช้ข้อมูลจำนวนหลายเรื่องและตัวแปรอยู่ในกลุ่มเดียวกันอาจจะให้ผลสอดคล้องกับงานวิจัยของณัฐวุฒิ ศิริกุลรุ่งโรจน์ และคณะ (2556) ก็มีความเป็นไปได้

5.3 ข้อเสนอแนะ

1. ตัวแปรที่นำมาใช้ในการงานวิจัยนี้เป็นเพียงส่วนหนึ่งของการพิจารณาการเป็นโรคเบาหวานเท่านั้น เพื่อให้การทำนายมีประสิทธิภาพมากขึ้น ควรเพิ่มตัวแปรที่เกี่ยวข้องอื่น ๆ อีก

2. เพื่อให้ผลสรุปครอบคลุมกว้างขวางเพิ่มขึ้น ควรจะทำการศึกษาวิธีอื่น ๆ ที่เป็นเทคนิคการทำเหมืองข้อมูลในการจำแนกกลุ่มเหมือนกัน เช่น วิธีฐานกฎ (Base-rule method) วิธีโครงข่ายความเชื่อของเบย์เซียน (Bayesian belief network method)

3. เพื่อให้ได้ข้อสรุปของผลการวิเคราะห์ข้อมูลที่มีความสมบูรณ์มากขึ้น เราอาจจะใช้อัลกอริทึมประเภทอื่น ๆ โดยวิธีความใกล้เคียงกันมากที่สุดยังมีอัลกอริทึม KStar และ LWL วิธีแผนภาพต้นไม้เพื่อการตัดสินใจมีอัลกอริทึม Decision Stump, LMT, Random Forest, Random Tree และ REP Tree วิธีโครงข่ายประสาทสามารถกำหนดค่าอัตราการเรียนรู้และค่าโมเมนตัมที่ละเอียดมากขึ้นกว่าเดิม อาจกำหนดจำนวนชั้นซ่อนมากกว่า 1 ชั้นได้ และอาจเพิ่มจำนวนรอบการสอนให้มากขึ้น เช่น 100,000 รอบ และวิธีซัพพอร์ตเวกเตอร์แมชชีนมีอัลกอริทึม Normalized Poly Kernel, REF Kernel และ Puk

4. นอกจากที่กล่าวมาแล้วในข้อที่ 3 เราอาจจะวิเคราะห์ข้อมูลด้วยอัลกอริทึมประเภทอื่น ๆ อีก วิธีซัพพอร์ตเวกเตอร์แมชชีนมี Kernel Function อีก 2 แบบ คือ Radial Basis Function Kernel และ Sigmoid Kernel และวิธีแผนภาพต้นไม้เพื่อการตัดสินใจมีอัลกอริทึมที่ใช้ในการจำแนกกลุ่มอื่น ๆ อีก เช่น ID3

ภาคผนวก ก

รายละเอียดและตัวอย่างข้อมูลที่นำมาวิเคราะห์

ข้อมูลของการพิจารณาการเป็นโรคเบาหวาน

ตารางที่ ก-1 คุณลักษณะและรายละเอียดของการพิจารณาการเป็นโรคเบาหวาน

คุณลักษณะ	รายละเอียด
1. เพศ (SEX)	ชาย (Male) หญิง (Female)
2. กรรมพันธุ์ (RELATIVE)	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 (EATTING)	หวาน (Sweet) มัน (Oily) เค็ม (Salt) ไม่ชอบ (No)
7. อุปนิสัยการรับประทานอาหาร (EATING BEHAVIOUR)	กินอาหารสุกๆดิบๆ (half-cook) ไม่กิน (No) นานๆครั้ง (Rarely) กินบ่อย (often)
8. อารมณ์ (TEMPERAMENT ON 1 MONTH AGO)	หดหู่ (Sad) เบื่อหน่าย (Boring) ไม่มีอาการดังกล่าว ๖No)
9. น้ำหนัก (WEIGHT)	หน่วยเป็นกิโลกรัม

คุณลักษณะ	รายละเอียด
10. ส่วนสูง (HEIGHT)	หน่วยเป็นเมตร
11. ค่าดัชนีมวลกาย (BMI)	เป็นตัวเลข (Numeric) คำนวณได้จาก $\text{BMI} = \frac{\text{WEIGHT}}{\text{HEIGHT}^2}$ หน่วย กก./ม^2
12. เกณฑ์ดัชนีมวลกาย (BMI MARK)	น้ำหนักน้อยเกินไป (Skinny) น้ำหนักปกติ (Thin) น้ำหนักเกิน (Shapely) อ้วนระดับ 1 (Plump) อ้วนระดับ 2 (Fat)
13. รอบเอว (WAISTLINE)	หน่วยเป็นเซนติเมตร
14. การเป็นโรคความดันโลหิต (RESULT)	เป็นโรคเบาหวาน (Positive) ไม่เป็นโรคเบาหวาน (Negative)

ตารางที่ ก-2 ตัวอย่างข้อมูลการพิจารณาการเป็นโรคเบาเหวาน

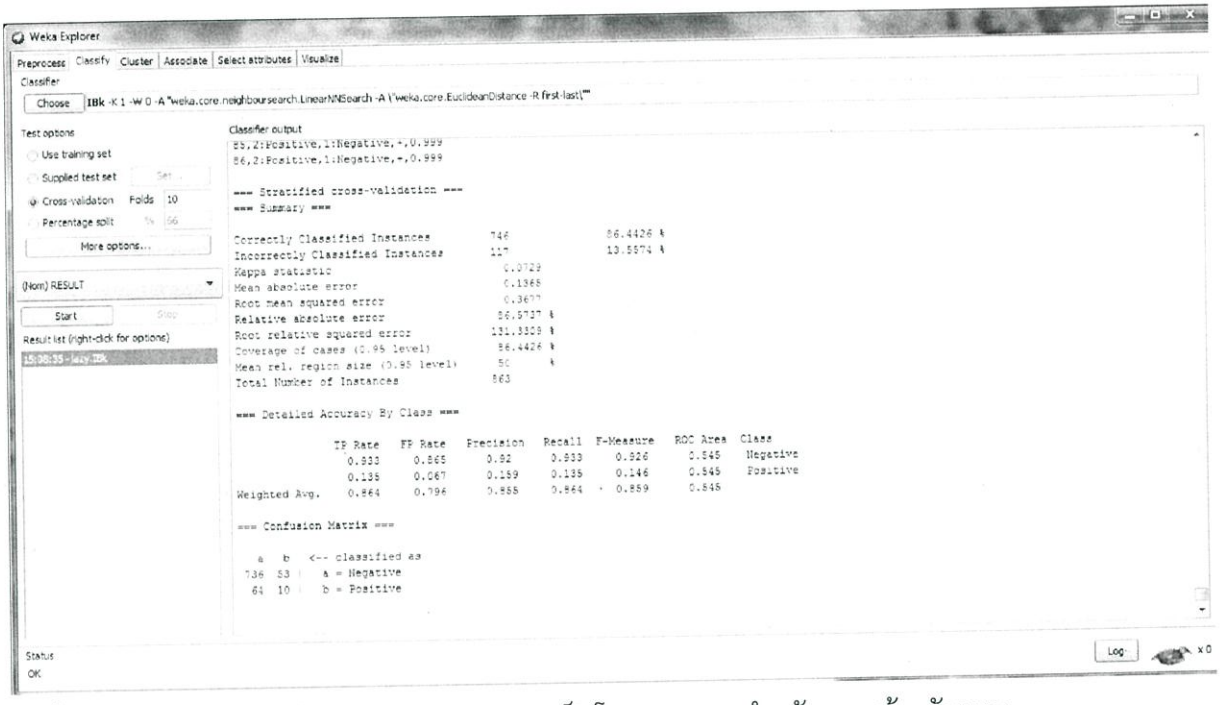
No	SEX	RELATIVE	SMOKE	ALCOHOL	EXERCISE	EATING	EATING BEHAVIOUR
1	Female	Unknow	No	No	Less	No	
2	Female	No	No	No	Everyday	Oily	No
3	Female	No	No	No	Everyday	Sweet	
4	Male	Unknow	No	Raraly	Equal	Salt	Rarely
5	Female	No	No	No	Everyday	No	No
6	Female	No	No	No	Everyday	No	No
7	Female	No	No	No	Everyday	No	No
8	Male	No	No	Yes	Everyday	No	No
9	Female	No	No	No	Everyday	Sweet	No
10	Female	Yes	No	No	Equal	Sweet	Rarely
11	Male	No	Yes	Ever	Everyday	No	No
12	Female	No	No	No	Everyday	Oily	Rarely
13	Female	No	Yes	Ever	Equal	Salt	Rarely
14	Male	No	No	Raraly	Equal	Sweet	Rarely
15	Female	Unknow	No	No	Everyday	Sweet	No
16	Female	Yes	No	No	Equal	Oily	No
17	Female	No	No	No	Everyday	Oily	Rarely
18	Male	No	No	Ever	Equal	No	No
19	Female	No	No	Ever	Equal	Sweet	Rarely
20	Female	Yes	No	Ever	Equal	Salt	
21	Female	No	No	No	Everyday	Sweet	Rarely
22	Female	No	No	No	Everyday	Oily	Rarely
23	Female	No	No	Raraly	Equal	No	
24	Male	No	No	No	Everyday	Sweet	No
25	Female	No	No	Raraly	Equal	No	Rarely
26	Female	Unknow	No	Ever	Equal	Salt	No
27	Female	Yes	No	Ever	Equal	No	No
28	Female	No	No	Ever	Equal	No	No

No	TEMPERAMENT ON 1 MONTH AGO	WEIGHT	HEIGH	BMI	BMI MARK	WAISTLINE	RESULT
1		55	1.58	22.0	Thin	81	Negative
2	No	54	1.5	24.0	Shapely	79	Negative
3	No	67	1.5	29.8	Plump	106	Negative
4	No	51	1.5	22.7	Shapely	97	Negative
5	No	70	1.56	28.8	Plump	109	Negative
6	No	59	1.5	26.2	Plump	88	Negative
7	No	50	1.56	20.5	Thin	76	Negative
8	No	47	1.76	15.2	Skinny	66	Negative
9	No	53	1.45	25.2	Plump	86	Negative
10	No	75	1.6	29.3	Plump	97	Negative
11	No	71	1.65	26.1	Plump	94	Negative
12	No	60	1.6	23.4	Shapely	92	Negative
13	No	64	1.65	23.5	Shapely	88	Negative
14	No	68	1.68	24.1	Shapely	91	Negative
15	No	65	1.58	26.0	Plump	90	Positive
16	No	55	1.65	20.2	Thin	87	Negative
17	No	66	1.6	25.8	Plump	92	Negative
18	No	48	1.6	18.8	Thin	76	Negative
19	No	66	1.6	25.8	Plump	93	Negative
20		46	1.6	18.0	Skinny	75	Negative
21	No	70	1.58	28.0	Plump	102	Negative
22	No	72	1.65	26.4	Plump	99	Negative
23		55	1.45	26.2	Plump	83	Negative
24	No	70	1.7	24.2	Shapely	98	Negative
25	No	46	1.45	21.9	Shapely	77	Negative
26	No	61	1.6	23.8	Shapely	91	Negative
27	No	52	1.5	23.1	Shapely	76	Negative
28	No	45	1.52	19.5	Thin	73	Negative

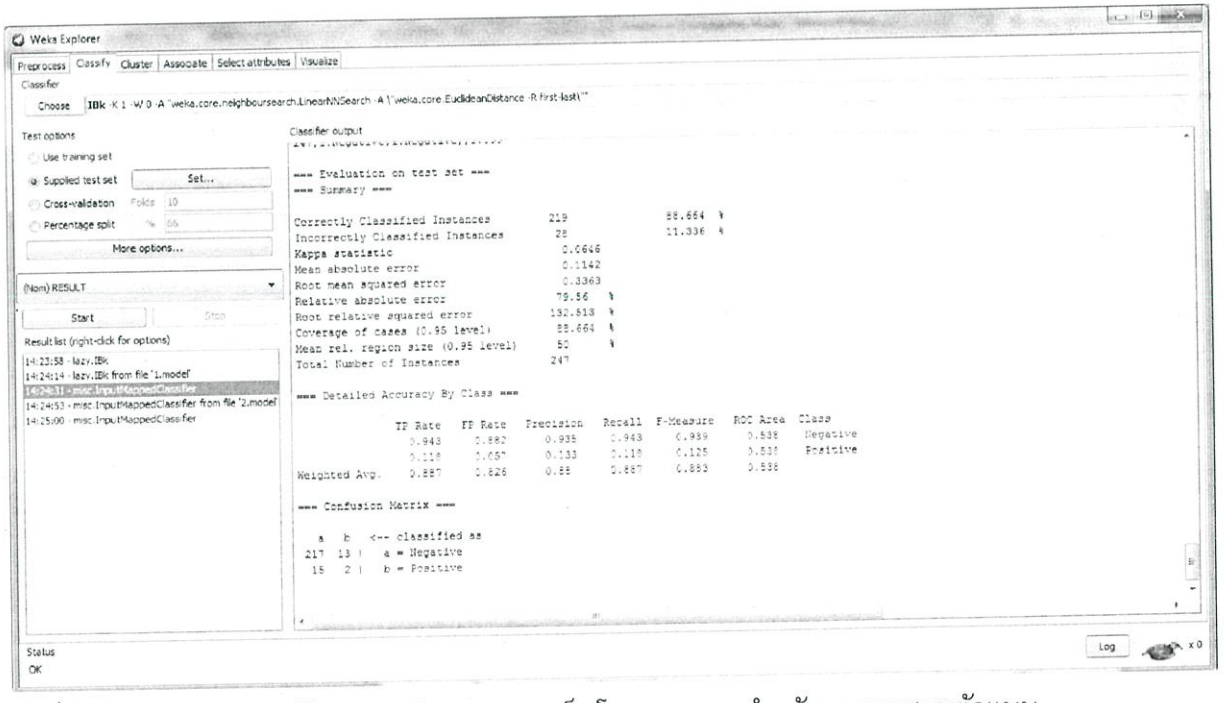
ภาคผนวก ข

การวิเคราะห์ข้อมูล

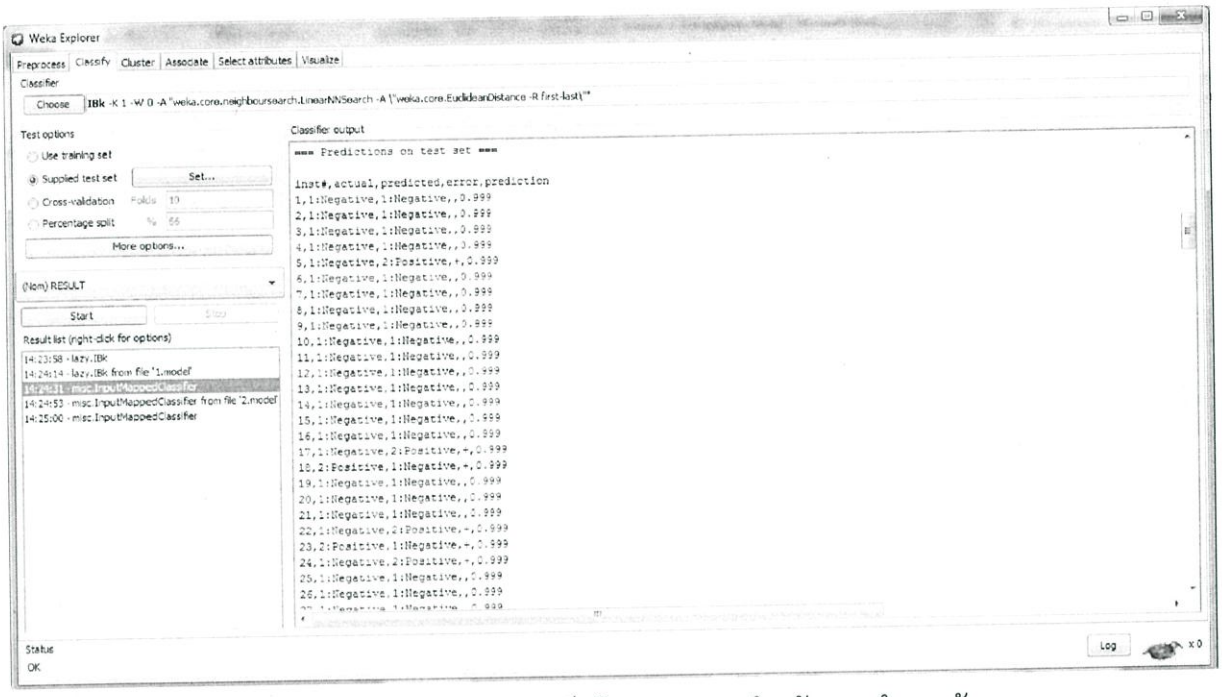
1. วิธีความใกล้เคียงกันมากที่สุด (K-Nearest Neighbor : KNN)



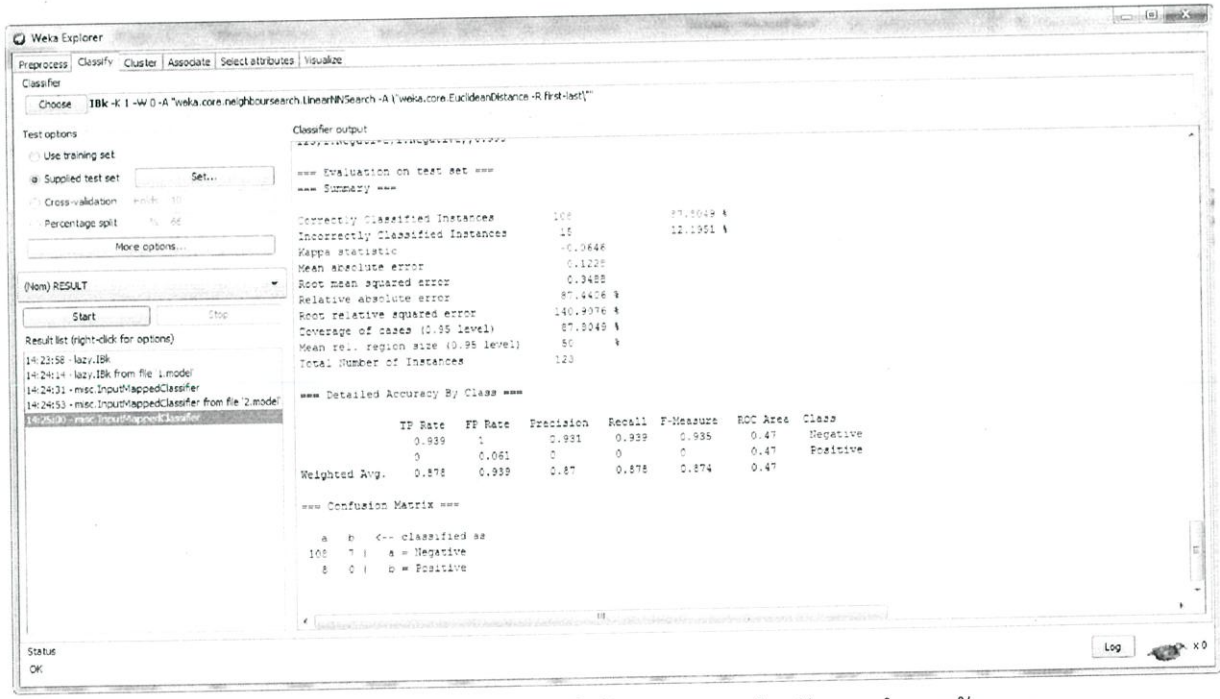
รูปที่ ข-1 ผลการวิเคราะห์ข้อมูลการพิจารณาการเป็นโรคเบาหวานสำหรับการสร้างตัวแบบโดยวิธีความใกล้เคียงกันมากที่สุด



รูปที่ ข-2 ผลการวิเคราะห์ข้อมูลการพิจารณาการเป็นโรคเบาหวานสำหรับการทดสอบตัวแบบโดยวิธีความใกล้เคียงกันมากที่สุด

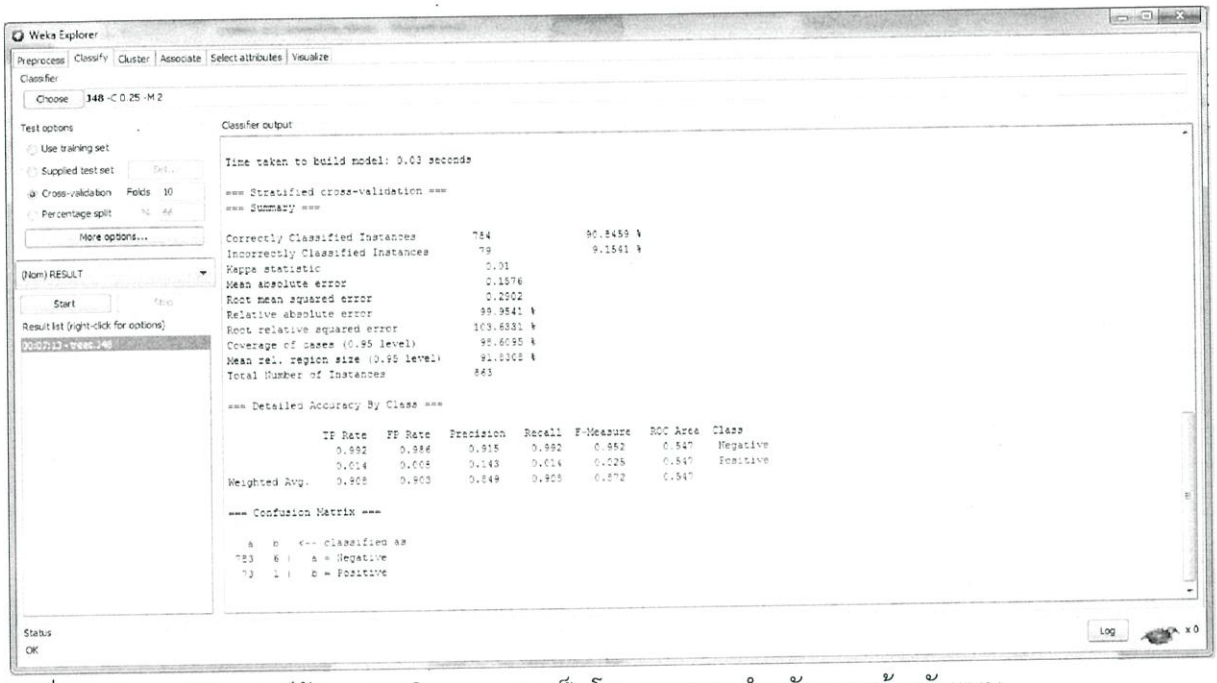


รูปที่ ข-3 ผลการวิเคราะห์ข้อมูลการพิจารณาการเป็นโรคเบาหวานสำหรับการทำนายตัวแบบโดยวิธีความใกล้เคียงกันมากที่สุด

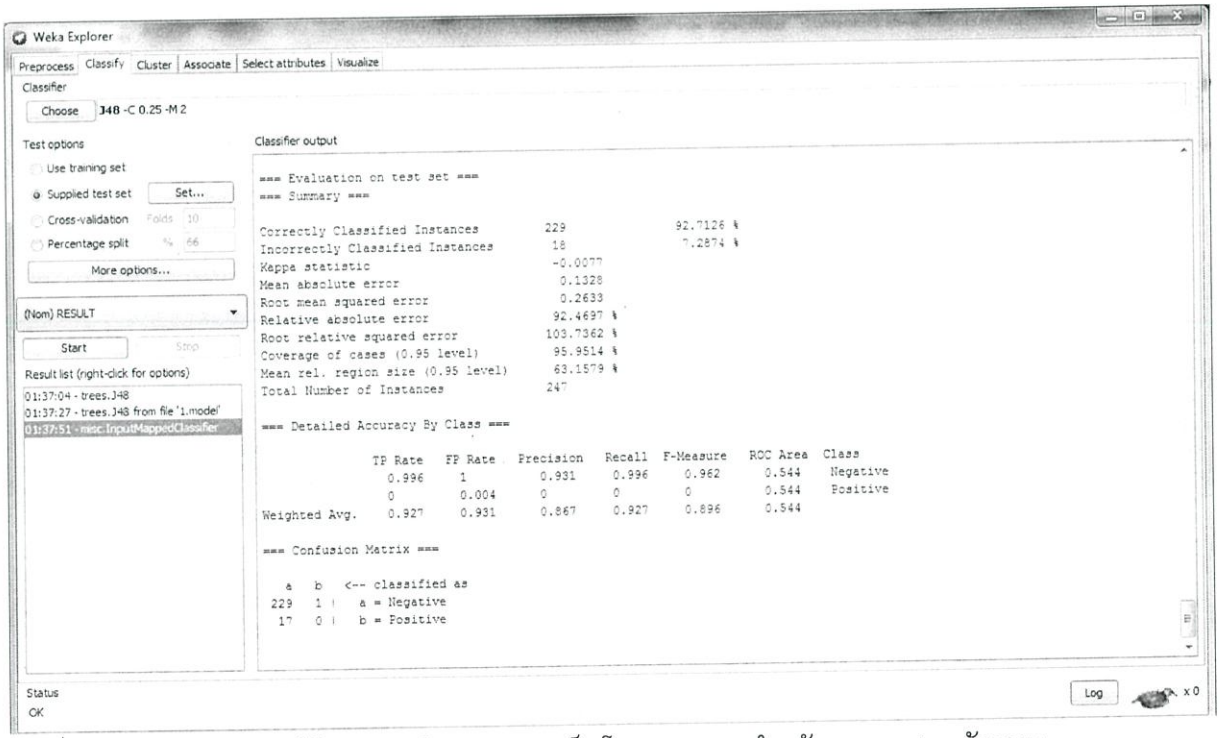


รูปที่ ข-4 ผลการวิเคราะห์ข้อมูลการพิจารณาการเป็นโรคเบาหวานสำหรับการทำนายตัวแบบโดยวิธีความใกล้เคียงกันมากที่สุด

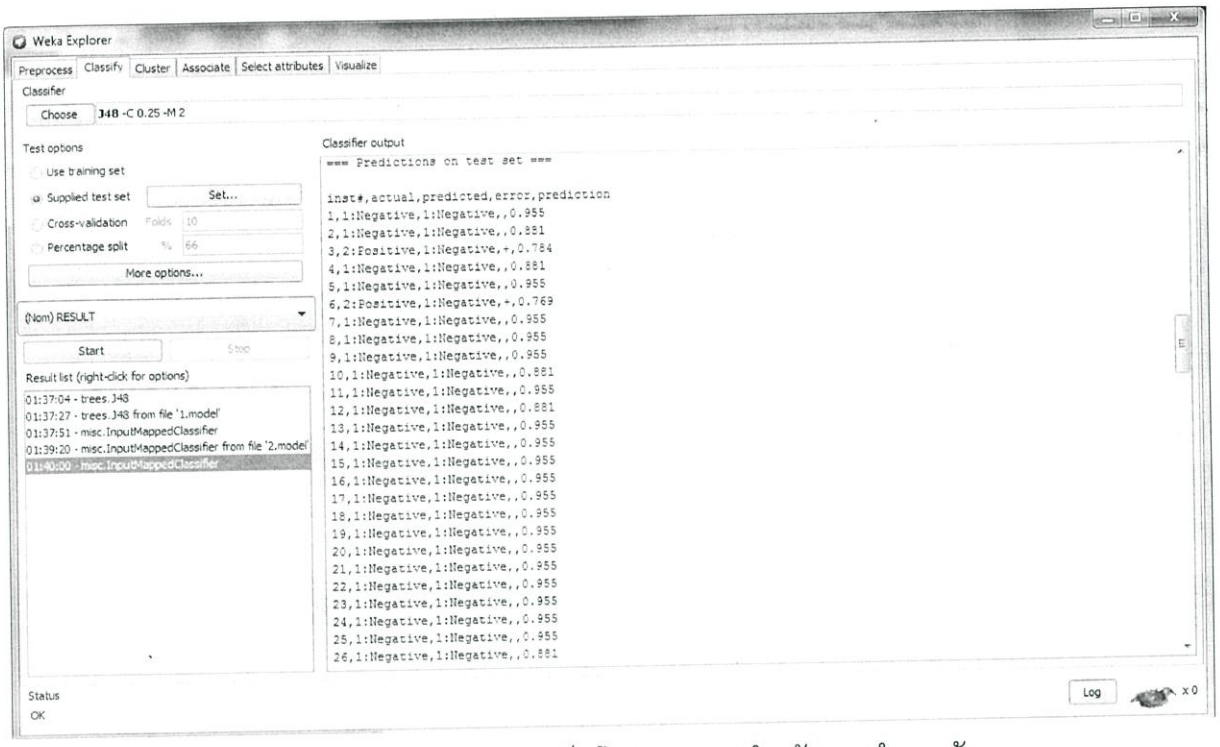
2. วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ (Decision Tree)



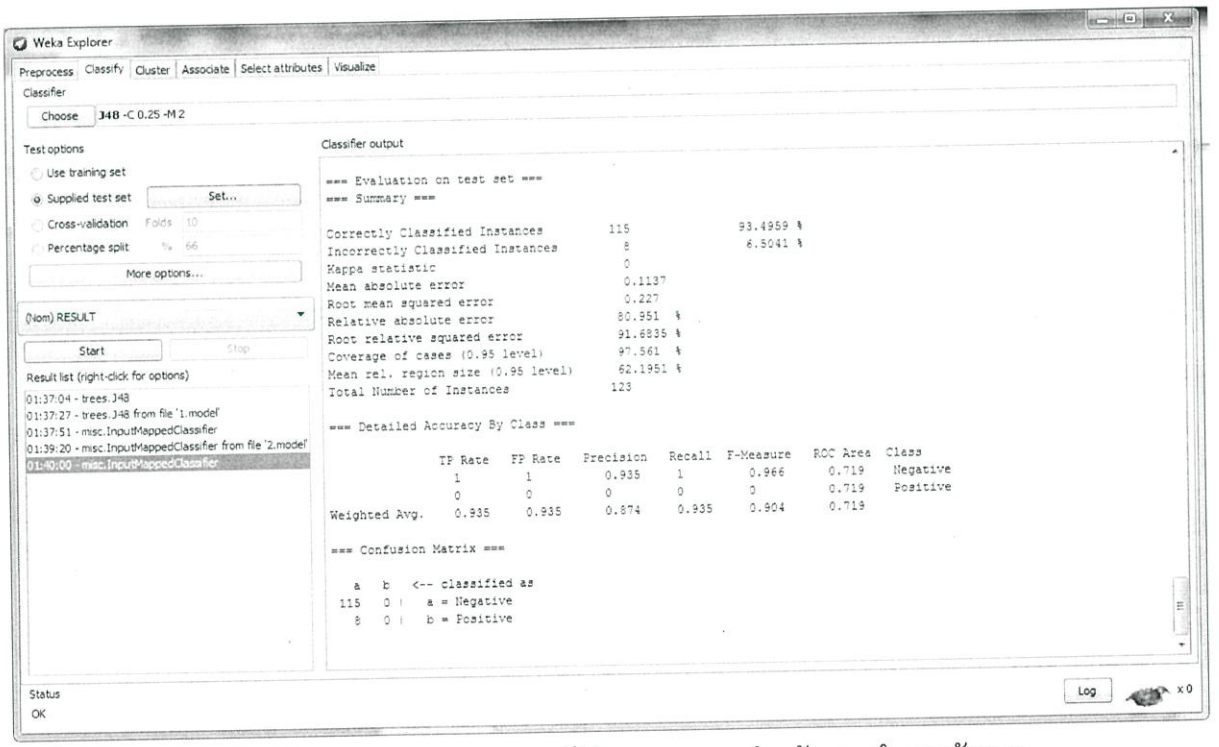
รูปที่ ข-5 ผลการวิเคราะห์ข้อมูลการพิจารณาการเป็นโรคเบาหวานสำหรับการสร้างตัวแบบโดยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ



รูปที่ ข-6 ผลการวิเคราะห์ข้อมูลการพิจารณาการเป็นโรคเบาหวานสำหรับการทดสอบตัวแบบโดยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ

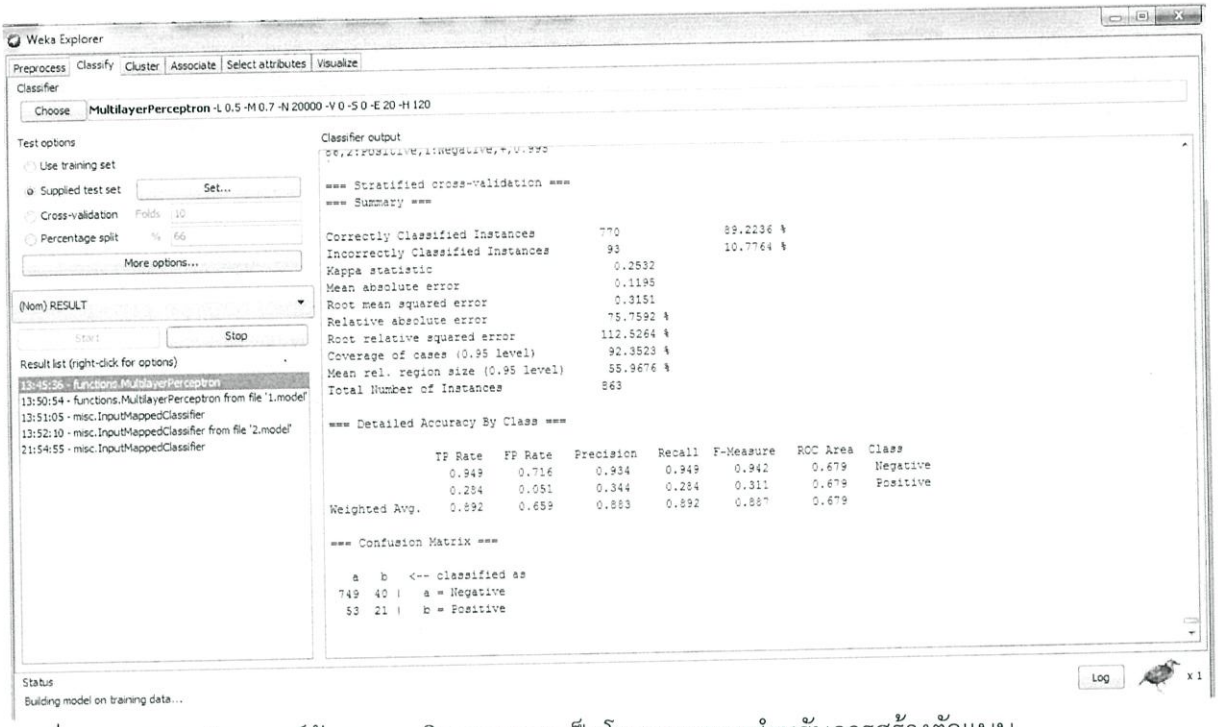


รูปที่ ข-7 ผลการวิเคราะห์ข้อมูลการพิจารณาการเป็นโรคเบาหวานสำหรับการทำนายตัวแบบโดยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ

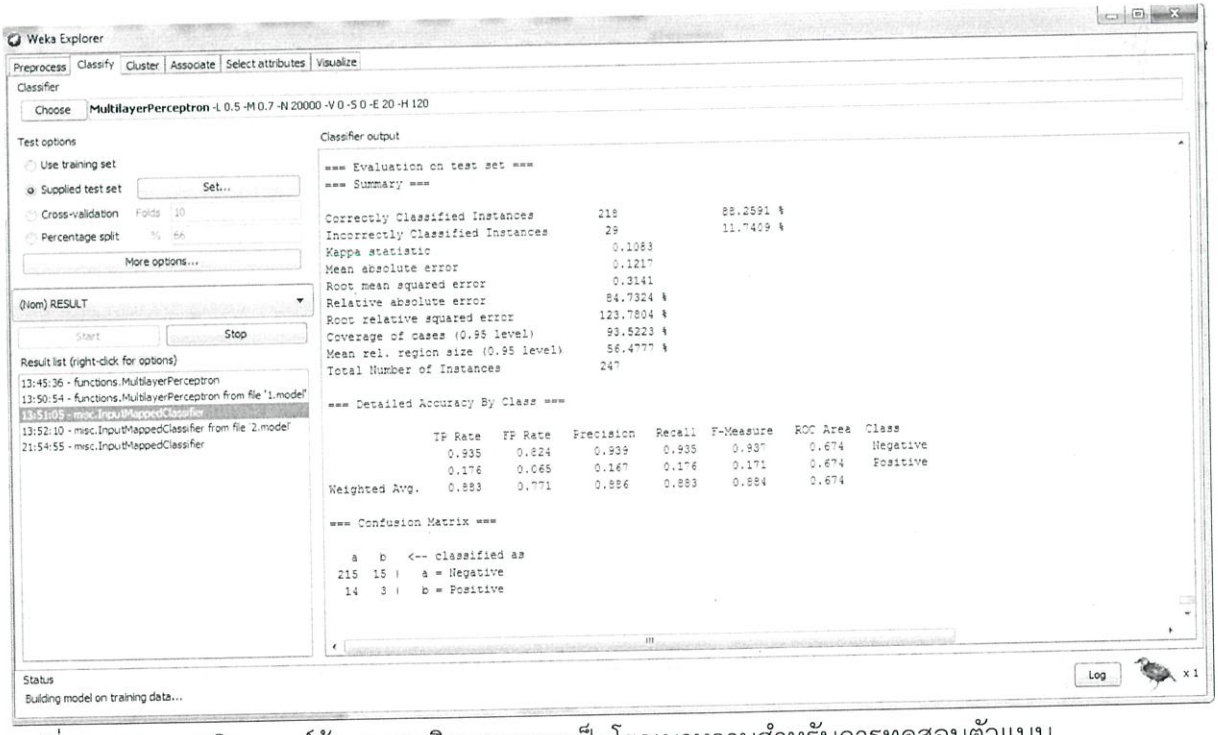


รูปที่ ข-8 ผลการวิเคราะห์ข้อมูลการพิจารณาการเป็นโรคเบาหวานสำหรับการทำนายตัวแบบโดยวิธีแผนภาพต้นไม้เพื่อการตัดสินใจ

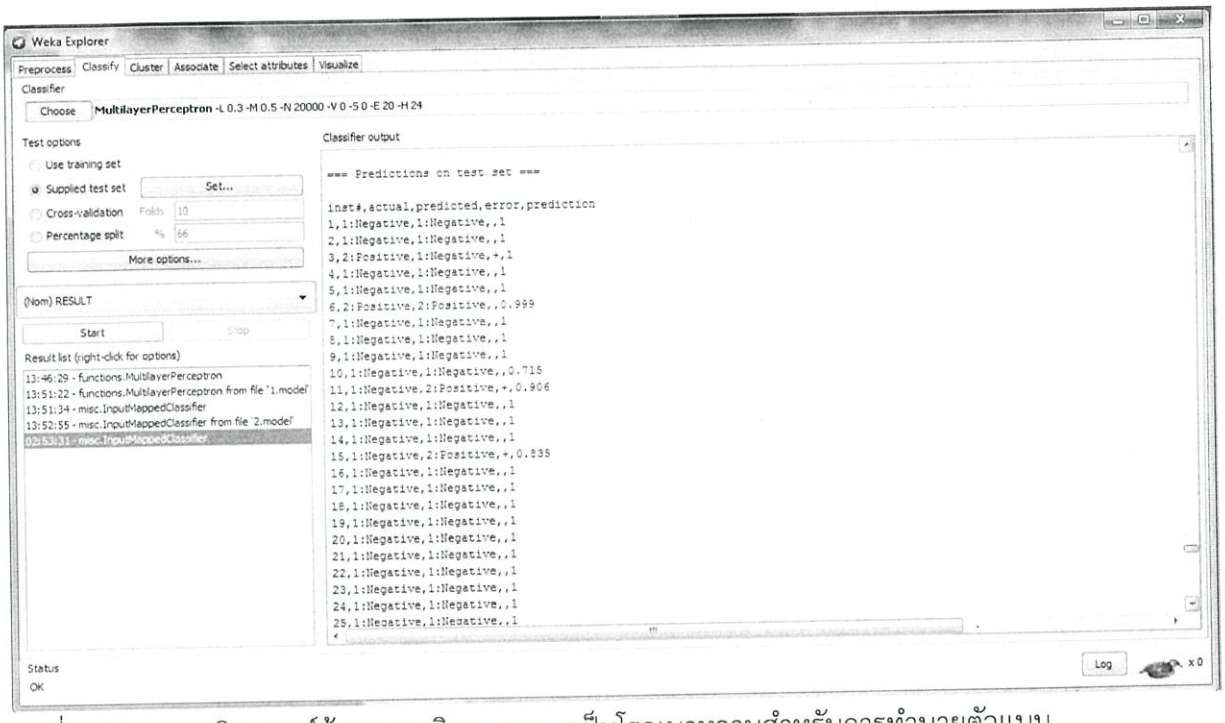
3. วิธีโครงข่ายประสาท (Neural Network)



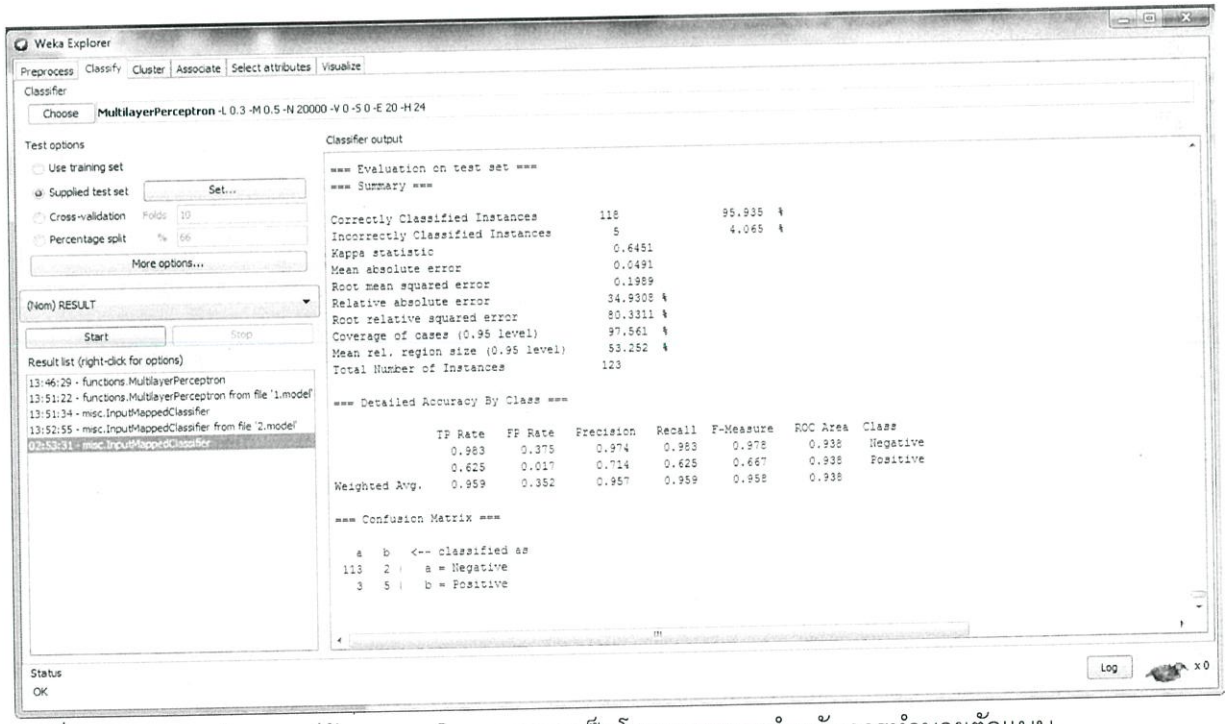
รูปที่ ข-9 ผลการวิเคราะห์ข้อมูลการพิจารณาการเป็นโรคเบาหวานสำหรับการสร้างตัวแบบโดยวิธีโครงข่ายประสาท



รูปที่ ข-10 ผลการวิเคราะห์ข้อมูลการพิจารณาการเป็นโรคเบาหวานสำหรับการทดสอบตัวแบบโดยวิธีโครงข่ายประสาท

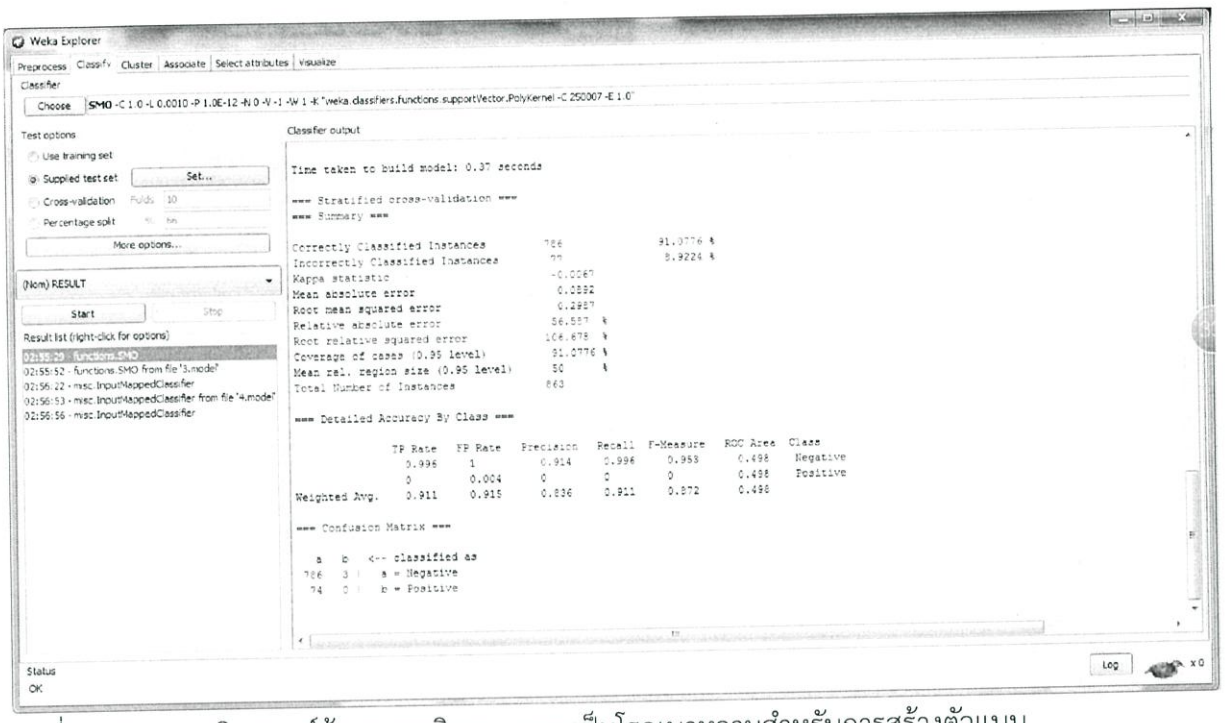


รูปที่ ข-11 ผลการวิเคราะห์ข้อมูลการพิจารณาการเป็นโรคเบาหวานสำหรับการทำนายตัวแบบโดยวิธีโครงข่ายประสาท

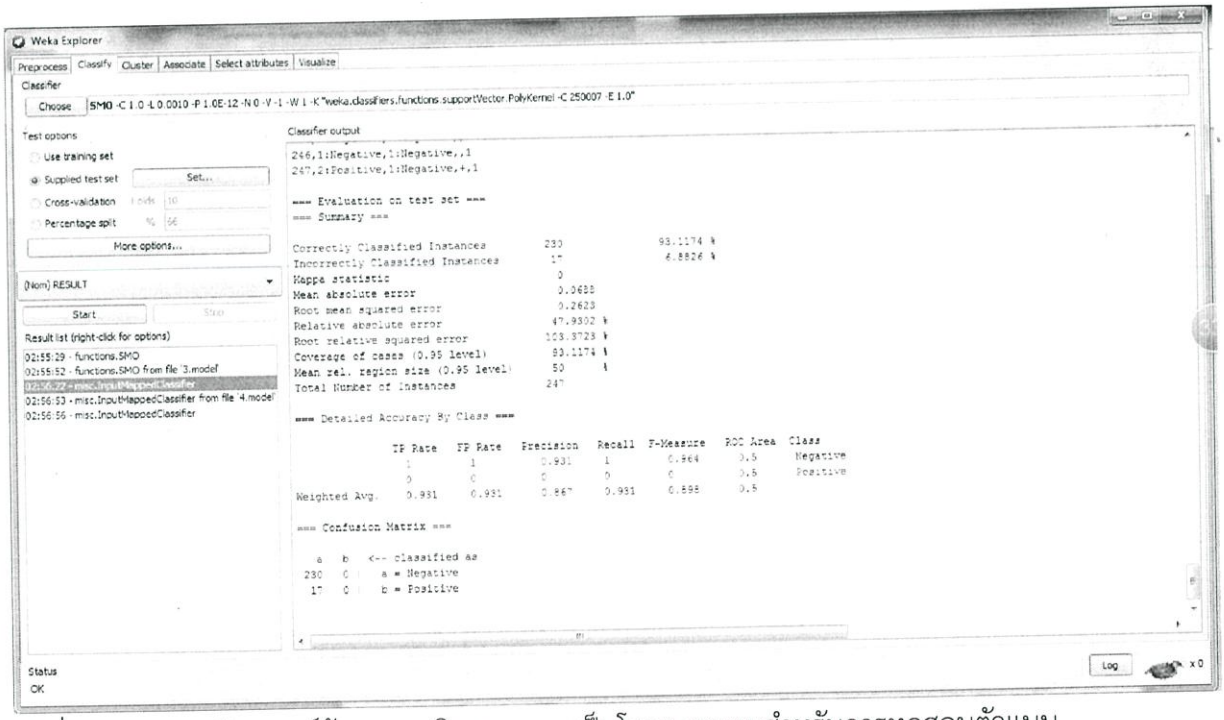


รูปที่ ข-12 ผลการวิเคราะห์ข้อมูลการพิจารณาการเป็นโรคเบาหวานสำหรับการทำนายตัวแบบโดยวิธีโครงข่ายประสาท

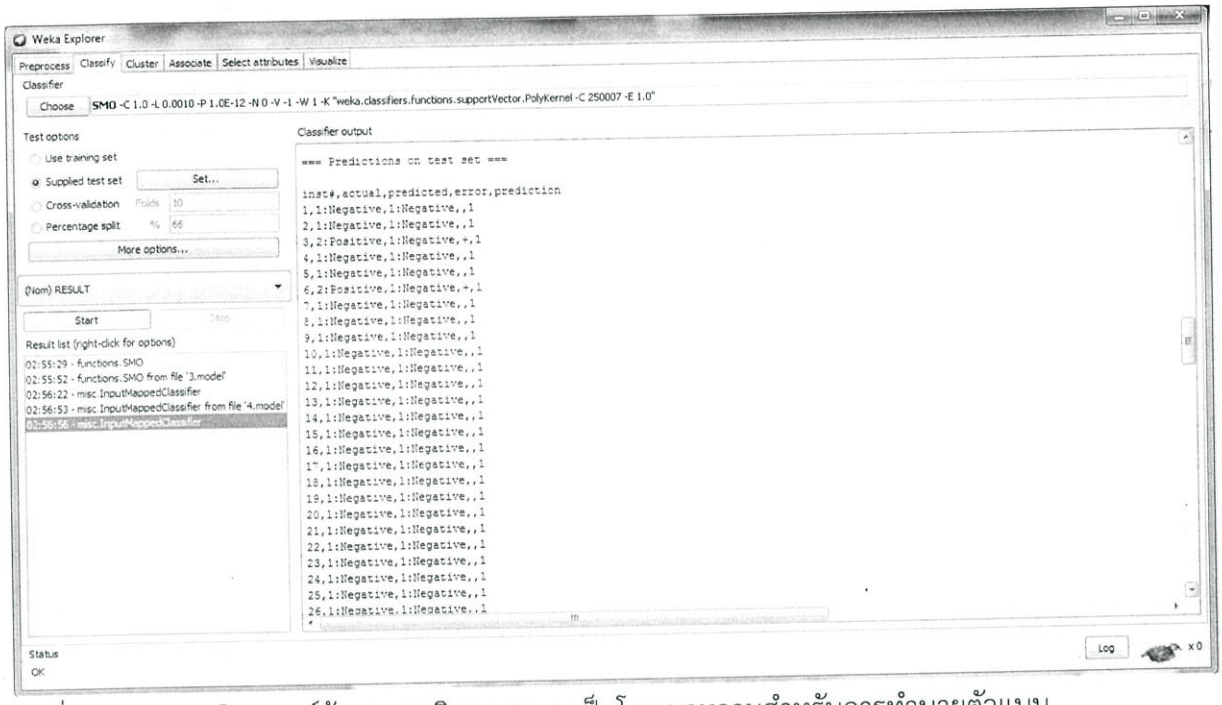
4. วิธีซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine)



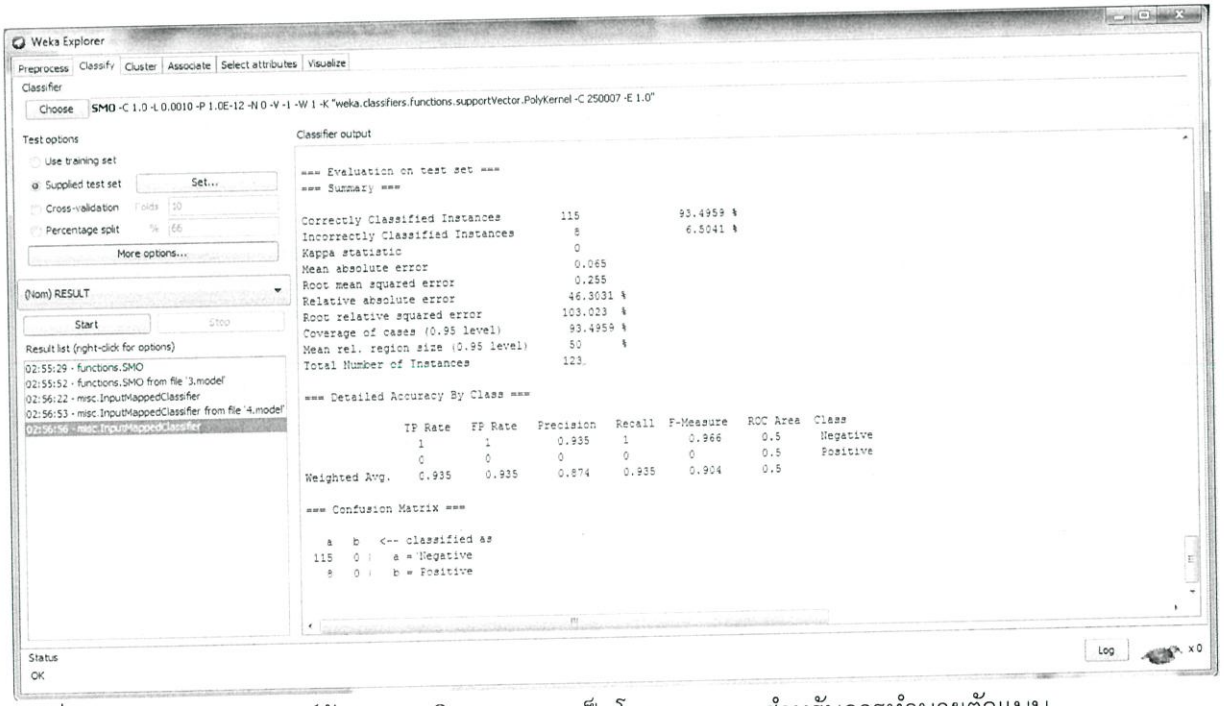
รูปที่ ข-13 ผลการวิเคราะห์ข้อมูลการพิจารณาการเป็นโรคเบาหวานสำหรับการสร้างตัวแบบ โดยวิธีซัพพอร์ทเวกเตอร์แมชชีน



รูปที่ ข-14 ผลการวิเคราะห์ข้อมูลการพิจารณาการเป็นโรคเบาหวานสำหรับการทดสอบตัวแบบ โดยวิธีซัพพอร์ทเวกเตอร์แมชชีน

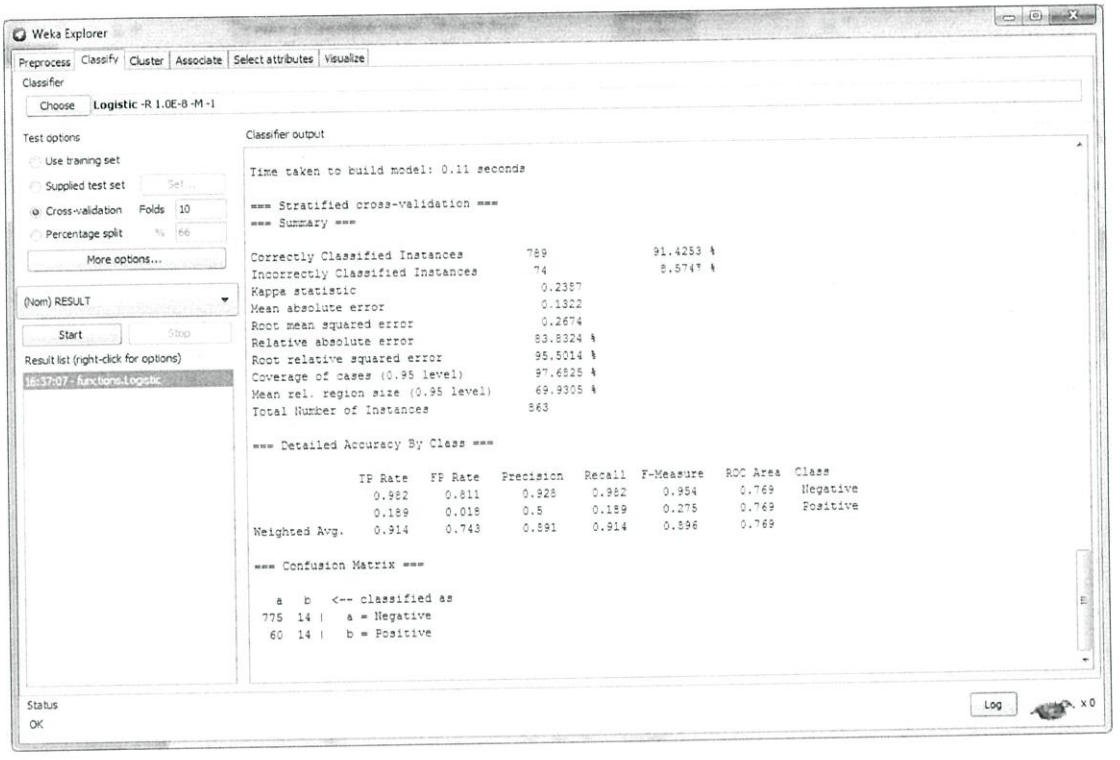


รูปที่ ข-15 ผลการวิเคราะห์ข้อมูลการพิจารณาการเป็นโรคเบาหวานสำหรับการทำนายตัวแบบโดยวิธีซัพพอร์ทเวกเตอร์แมชชีน

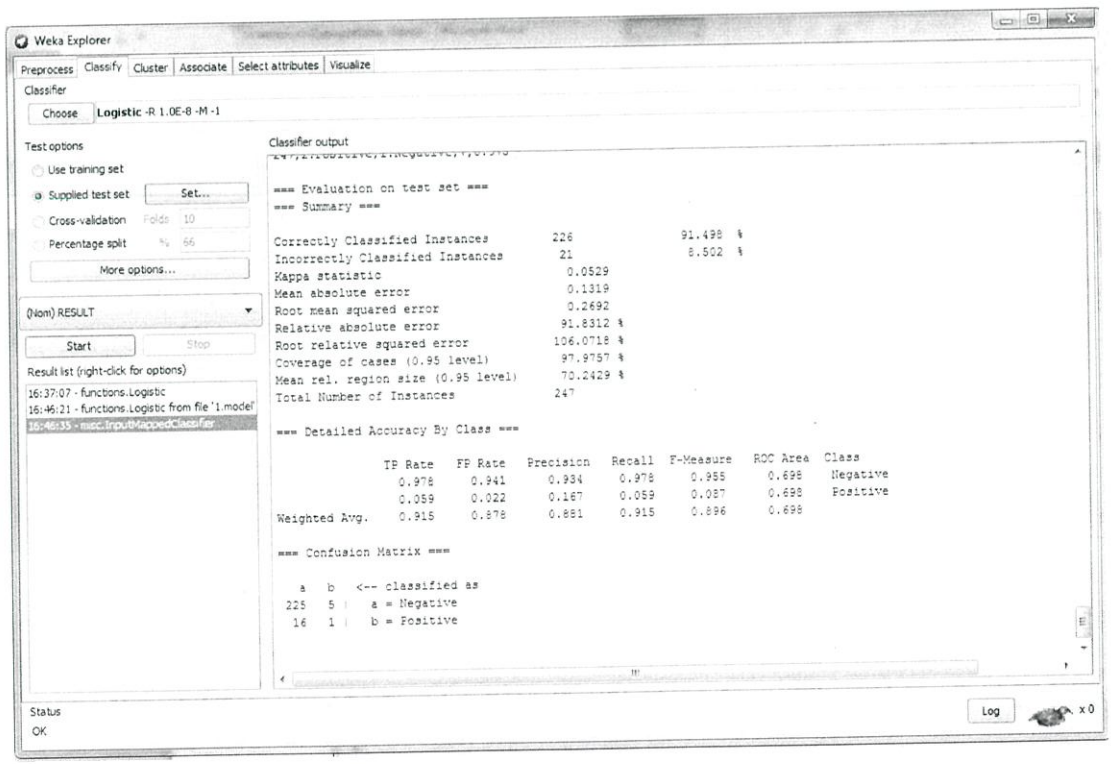


รูปที่ ข-16 ผลการวิเคราะห์ข้อมูลการพิจารณาการเป็นโรคเบาหวานสำหรับการทำนายตัวแบบโดยวิธีซัพพอร์ทเวกเตอร์แมชชีน

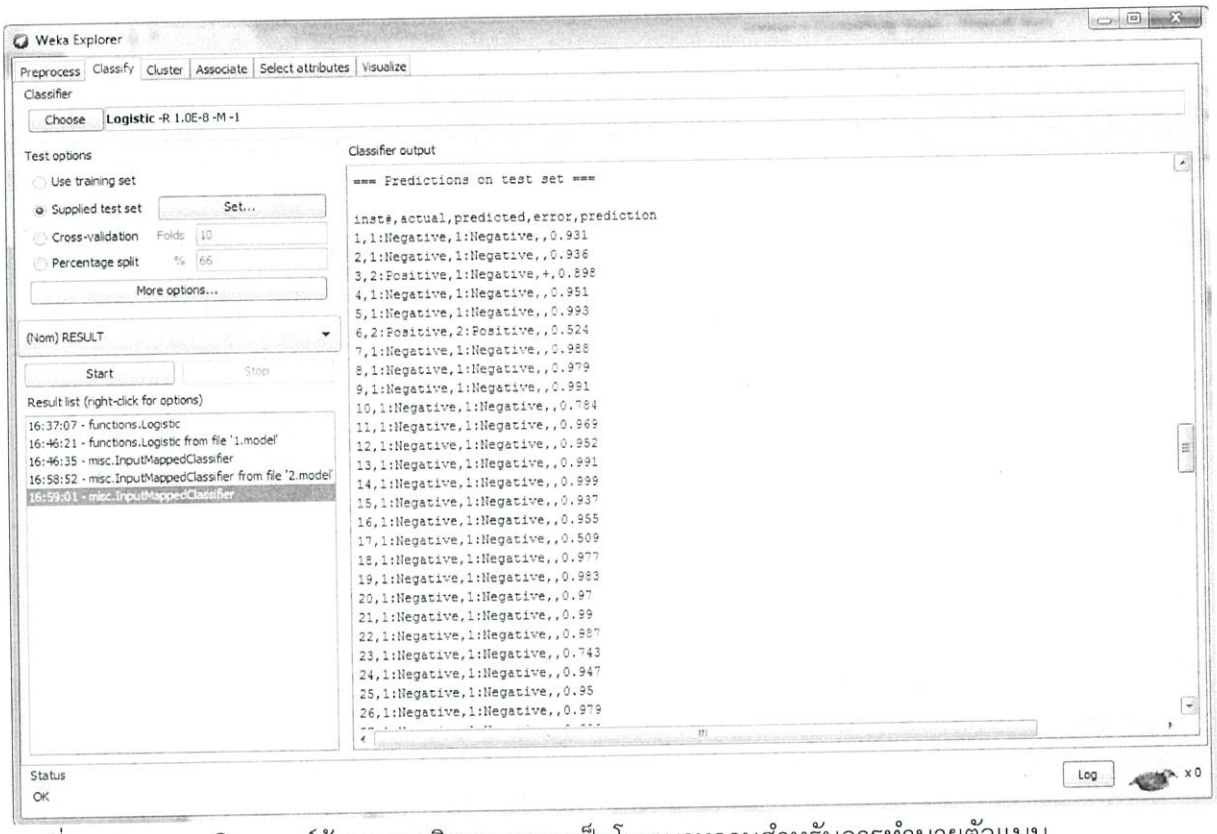
5. วิธีการถดถอยโลจิสติก (Logistic Regression)



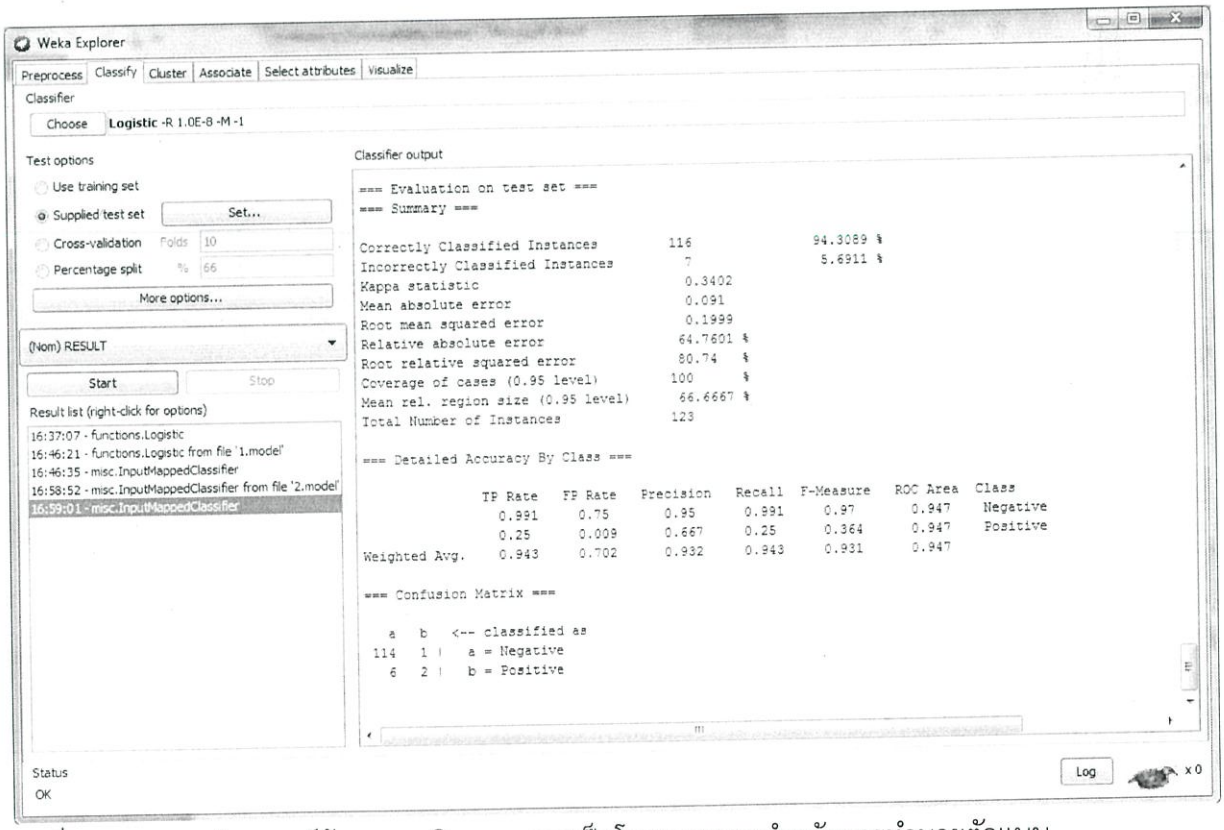
รูปที่ ข-17 ผลการวิเคราะห์ข้อมูลการพิจารณาการเป็นโรคเบาหวานสำหรับการสร้างตัวแบบโดยวิธีการถดถอยโลจิสติก



รูปที่ ข-18 ผลการวิเคราะห์ข้อมูลการพิจารณาการเป็นโรคเบาหวานสำหรับการทดสอบตัวแบบโดยวิธีการถดถอยโลจิสติก

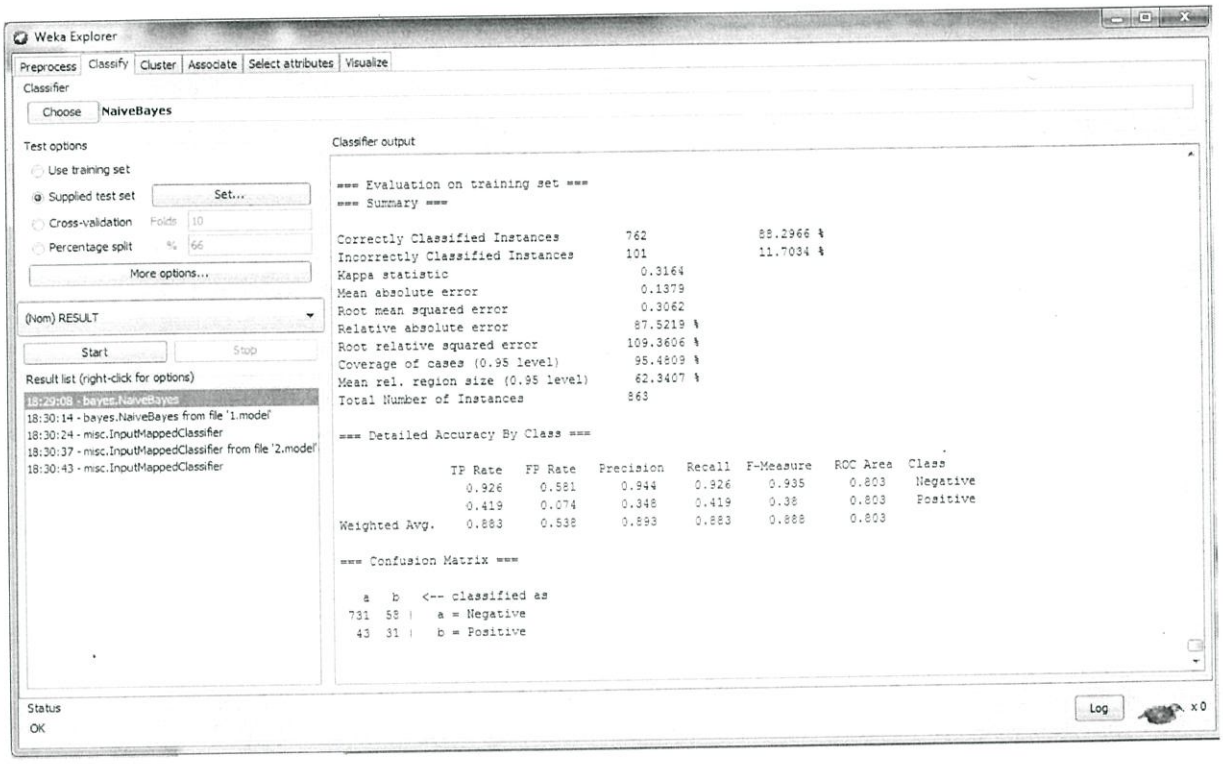


รูปที่ ข-19 ผลการวิเคราะห์ข้อมูลการพิจารณาการเป็นโรคเบาหวานสำหรับการทำนายตัวแบบโดยวิธีการถดถอยโลจิสติก

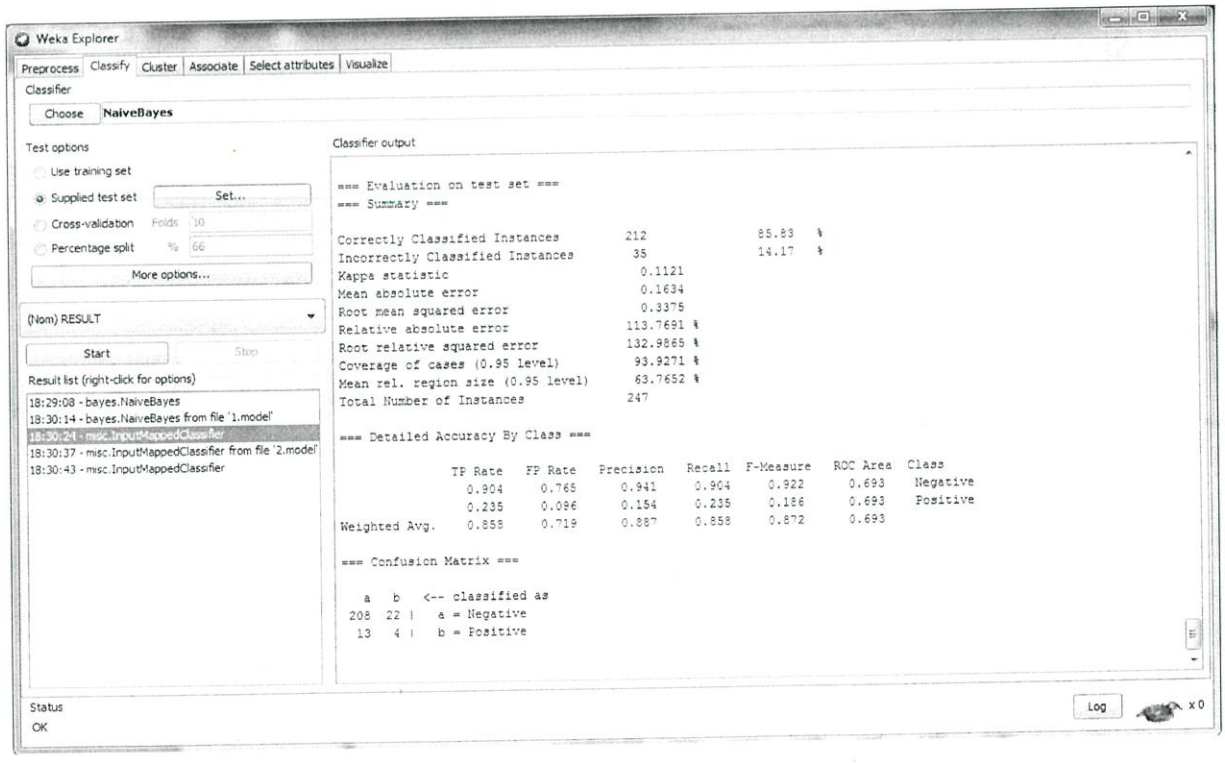


รูปที่ ข-20 ผลการวิเคราะห์ข้อมูลการพิจารณาการเป็นโรคเบาหวานสำหรับการทำนายตัวแบบโดยวิธีการถดถอยโลจิสติก

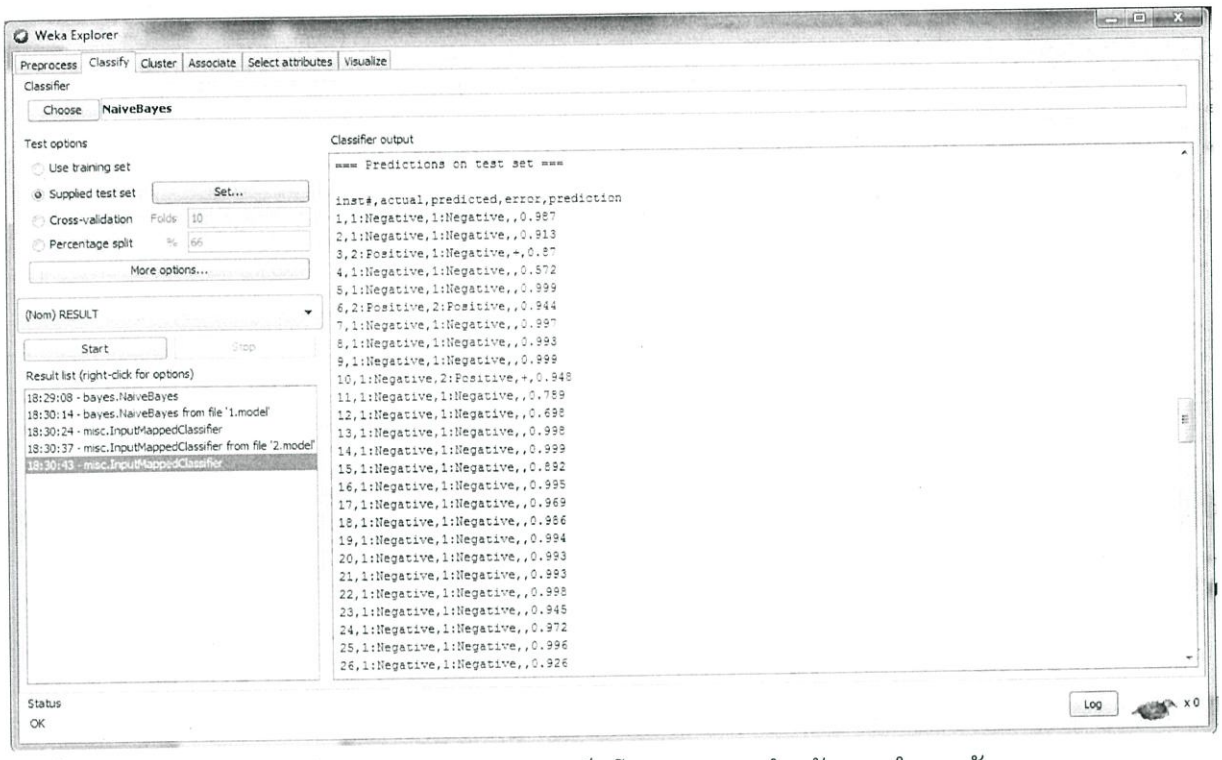
6. วิธีนาอีฟ เบย์ (Naïve Bayes)



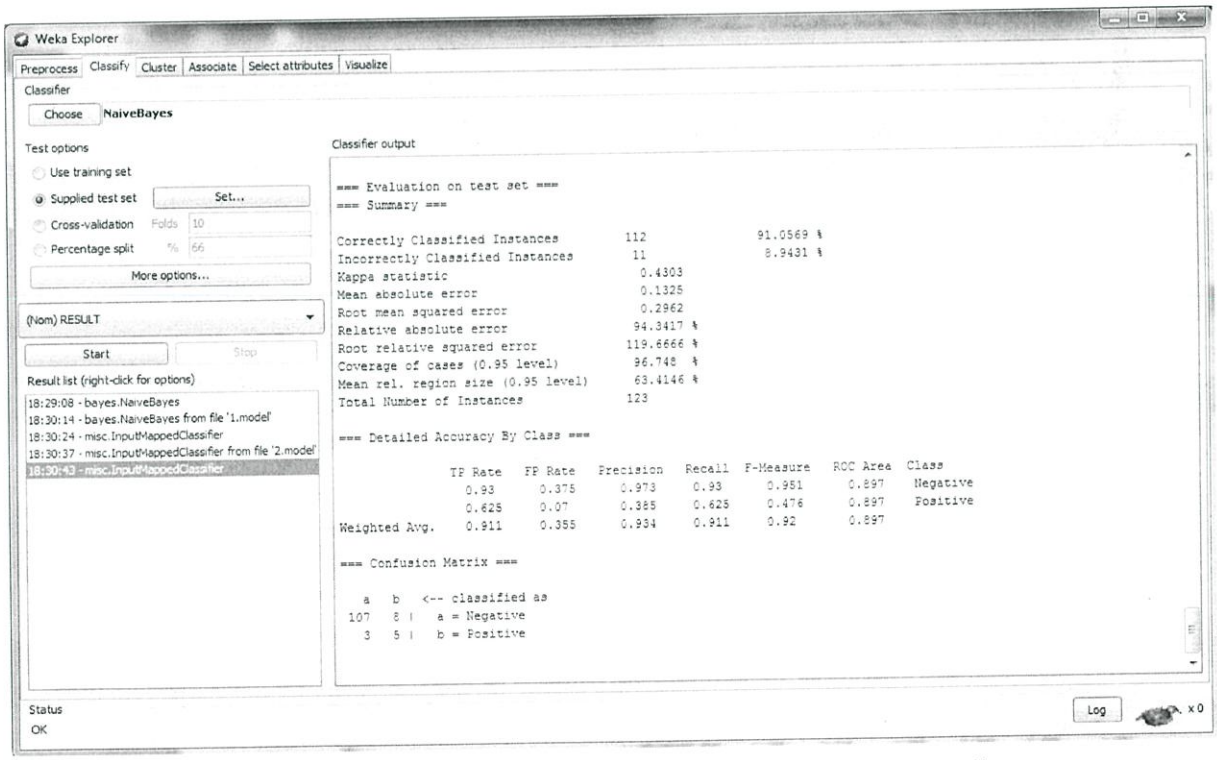
รูปที่ ข-17 ผลการวิเคราะห์ข้อมูลการพิจารณาการเป็นโรคเบาหวานสำหรับการสร้างตัวแบบโดยวิธีนาอีฟ เบย์



รูปที่ ข-18 ผลการวิเคราะห์ข้อมูลการพิจารณาการเป็นโรคเบาหวานสำหรับการทดสอบตัวแบบโดยวิธีนาอีฟ เบย์



รูปที่ ข-19 ผลการวิเคราะห์ข้อมูลการพิจารณาการเป็นโรคเบาหวานสำหรับการทำนายตัวแบบโดยวิธีนาอีฟ เบย์



รูปที่ ข-20 ผลการวิเคราะห์ข้อมูลการพิจารณาการเป็นโรคเบาหวานสำหรับการทำนายตัวแบบโดยวิธีนาอีฟ เบย์

ภาคผนวก ค

ตัวอย่างการคำนวณ

ตัวอย่างที่ 1 การคำนวณค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) ค่าความถ่วงดุล (F-Measure) ของการวิเคราะห์ข้อมูลการพิจารณาการเป็นโรคเบาหวาน โดยวิธีโครงข่ายประสาท

จากรูปที่ ข-12

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

$= \frac{113+5}{113+5+3+2}$

= 0.9593 หรือ 95.935%

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

$= \frac{113}{113+3}$

= 0.9741 หรือ 97.41%

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

$= \frac{113}{113+2}$

= 0.9826 หรือ 98.26%

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

$= \frac{2\times(0.9826\times 0.9741)}{0.9826+0.9741}$

= 0.9783 หรือ 97.83%

ตัวอย่างที่ 2 การคำนวณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) ของการวิเคราะห์ข้อมูลการพิจารณาการเป็นโรคเบาหวานสำหรับการทำนายตัวแบบโดยวิธีโครงข่ายประสาท

จากรูปที่ ข-11

กำหนดให้ค่า $y_i = 1$ ได้จากกรณีที่ค่าจริง (actual) ใน class attribute ของข้อมูลการพิจารณาการเป็นโรคเบาหวานในระเบียนนั้นตรงกันกับค่าทำนาย (predicted)

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

จะได้ $y_i = y_2 = 1$

ระเบียนที่ 6 ค่าจริง (actual) = 2 : Positive ค่าทำนาย (predicted) = 2 : Positive

จะได้ $y_i = y_6 = 1$

กำหนดให้ค่า $y_i = 0$ ได้จากกรณีที่ค่าจริง (actual) ใน class attribute ของข้อมูลการพิจารณาการเป็นโรคเบาหวานในระเบียนนั้นไม่ตรงกันกับค่าทำนาย (predicted)

เช่น ระเบียนที่ 3 ค่าจริง (actual) = 2 : Positive ค่าทำนาย (predicted) = 1 : Negative

จะได้ $y_i = y_3 = 0$

ระเบียนที่ 11 ค่าจริง (actual) = 1 : Negative ค่าทำนาย (predicted) = 2 : Positive

จะได้ $y_i = y_{11} = 0$

กำหนดให้ค่า $\hat{y}_i$ ได้จากค่าการทำนาย (predicted) ซึ่งอยู่ที่คอลัมน์ขวาสุดของระเบียนนั้นในช่อง Classifier output

ลำดับที่	y_i	$\hat{y}_i$	e_i^2
1	1	1	0
2	1	1	0
3	0	1	1
4	1	1	0
5	1	1	0
6	1	0.999	0
7	1	1	0
8	1	1	0
9	1	1	0
10	1	0.715	0.081
11	0	0.906	0.821
12	1	1	0

ลำดับที่	y_i	$\hat{y}_i$	e_i^2
13	1	1	0
14	1	1	0
15	0	0.835	0.697
16	1	1	0
17	1	1	0
18	1	1	0
19	1	1	0
20	1	1	0
21	1	1	0
22	1	1	0
23	1	1	0
24	1	1	0
25	1	1	0
26	1	1	0
27	1	1	0
28	1	1	0
29	1	1	0
30	1	0.571	0.184
31	1	1	0
32	1	1	0
33	1	1	0
34	1	1	0
35	1	1	0
36	1	1	0
37	1	1	0
38	1	1	0
39	1	1	0
40	1	1	0
41	1	1	0
42	1	1	0
43	1	1	0

ลำดับที่	y_i	$\hat{y}_i$	e_i^2
44	1	1	0
45	1	1	0
46	1	1	0
47	1	0.759	0.058
48	1	1	0
49	1	1	0
50	1	0.999	0
51	1	1	0
52	1	1	0
53	1	1	0
54	1	1	0
55	1	1	0
56	1	1	0
57	1	1	0
58	1	1	0
59	1	1	0
60	1	1	0
61	0	1	1
62	1	1	0
63	1	1	0
64	1	1	0
65	1	0.884	0.013
66	1	1	0
67	1	1	0
68	1	1	0
69	0	1	1
70	1	1	0
71	1	1	0
72	1	1	0
73	1	1	0
74	1	0.934	0.004

ลำดับที่	y_i	$\hat{y}_i$	e_i^2
75	1	1	0
76	1	1	0
77	1	1	0
78	1	1	0
79	1	1	0
80	1	1	0
81	1	1	0
82	1	0.976	0.001
83	1	1	0
84	1	1	0
85	1	1	0
86	1	1	0
87	1	1	0
88	1	1	0
89	1	1	0
90	1	1	0
91	1	1	0
92	1	1	0
93	1	1	0
94	1	1	0
95	1	0.966	0.001
96	1	0.999	0
97	1	0.999	0
98	1	1	0
99	1	1	0
100	1	1	0
101	1	1	0
102	1	0.959	0.002
103	1	1	0
104	1	1	0
105	1	1	0

ลำดับที่	y_i	$\hat{y}_i$	e_i^2
106	1	1	0
107	1	1	0
108	1	1	0
109	1	1	0
110	1	1	0
111	1	1	0
112	1	1	0
113	1	1	0
114	1	1	0
115	1	1	0
116	1	1	0
117	1	1	0
118	1	1	0
119	1	1	0
120	1	1	0
121	1	1	0
122	1	1	0
123	1	0.946	0.003
รวม			**4.866**

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

$$= \frac{4.866}{123}$$

$$= 0.0396$$

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

$$= \sqrt{0.0396}$$

$$= 0.1989$$

เอกสารอ้างอิง

1) นุชรี อาบสุวรรณ และนิตยา พันธุเวทย์. 2554 สำนักโรคไม่ติดต่อ กรมควบคุมโรค กระทรวงสาธารณสุข.

2) http:/www.ifd.org/worlddiabetesday/act-on-diabetes-now

3) อมรา ทองหงส์. 2555. สำนักระบาดวิทยา กรมควบคุมโรค กระทรวงสาธารณสุข.

4) ณัฐวรรณ รัตนากรกุล. 2545. ระบบจำแนกประเภทโครงสร้างของโปรตีนโดยใช้เทคนิคดาต้าไมน์นิ่ง. งานประชุมวิชาการ มก. ครั้งที่ 40.

5) กิตติพล วิแสง สิรภัทร เชี่ยวชาญวัฒนา และคำรณ สุนัติ. 2552. การวิเคราะห์ปัจจัยเสี่ยงของโรคเบาหวาน. ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น. The 5th National Conference on Computing and Information Technology. p. 798-805.

6) ภัทรพงศ์ พงศ์ภัทรกานต์. 2552. การเปรียบเทียบการจำแนกข้อมูลของแบบจำลอง CART, SVM, C5.0 และแบบผสมผสานกัน. ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏเลย.

7) ภัทราวุฒิ แสงศิริ ศรีมาจ ณ วิเชียร และพยุง มีสัจ. 2553. การคัดแยกประเภทมะเร็งเม็ดเลือดขาวโดยใช้วิธีการจัดอันดับร่วมกับเทคนิคซัพพอร์ตเวกเตอร์แมชชีน. การประชุมทางวิชาการเสนอผลงานวิจัยระดับบัณฑิตศึกษา ครั้งที่ 11.

8) นิเวศ จิระวิชิตชัย. 2553. การค้นหาเทคนิคการทำเหมืองข้อมูลเพื่อสร้างโมเดลการวิเคราะห์โรคอัตโนมัติ. มหาวิทยาลัยราชภัฏสวนสุนันทา.

9) วาทินี นุ้ยเพียร และคณะ. 2553. การเปรียบเทียบประสิทธิภาพและวิเคราะห์การจำแนกข้อมูลโดยใช้โครงข่ายประสาทเทียม ซัพพอร์ตเวกเตอร์แมชชีน นาอีฟเบย์ และเคเนียรเรสต์เนเบอร์. ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ.

10) ชาญชัย พรมโคตร ทิพวรรณ ศรีชนะและอัจฉรา มหาวีรวัฒน์. 2554. การวิเคราะห์ข้อมูลความเสี่ยงของการเป็นโรคหอบหืดด้วยเทคนิคโครงข่ายประสาทเทียมแบบหลายชั้นร่วมกับขั้นตอนวิธีการเรียนรู้แบบแพร่ย้อนกลับ. ภาควิชาคณิตศาสตร์ สถิติ และคอมพิวเตอร์, คณะวิทยาศาสตร์, มหาวิทยาลัยอุบลราชธานี. The 5th UBU Conference Proceeding, 4-5 August 2011. p 66-72.

11) สุมนพิชญ์ พลศรี และอัจฉรา มหาวีรวัฒน์. 2554. การวิเคราะห์ความเสี่ยงของการเป็นวัณโรคด้วยเทคนิคโครงข่าย ประสาทเทียม. ภาควิชาคณิตศาสตร์ สถิติ และคอมพิวเตอร์, คณะวิทยาศาสตร์, มหาวิทยาลัยอุบลราชธานี. The 5th UBU Conference Proceeding, 4-5 August 2011. p 12-18.

เอกสารอ้างอิง (ต่อ)

12) เดช ธรรมศิริ และพยุง มีสัจ. 2554. การเรียนรู้แบบรวมกลุ่มด้วยโครงข่ายประสาทเทียมเอดาบูท สำหรับการจำแนกข้อมูล. ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ และภาควิชาครุศาสตร์ไฟฟ้า คณะครุศาสตร์อุตสาหกรรม มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ. วารสารเทคโนโลยีสารสนเทศ ปีที่ 7 ฉบับที่ 14 กรกฎาคม – ธันวาคม. หน้า 7-12.

13) นงเยาว์ ในอรุณ และพรรณี สิทธิเดช. 2555. การใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีนและโครงข่ายประสาทเทียมแบบแพร่กลับสำหรับการจำแนกผู้ป่วยโรคหัวใจขาดเลือดและโรคหัวใจรูปแบบอื่น. ภาควิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยนเรศวร จังหวัดพิษณุโลก. The 4th Science Research Conference. 12-13 March 2012. Faculty of Science, Naresuan University.

14) เพิ่มพันธุ์ ธรรมสโรช และพรภัทร ธรรมสโรช. 2555. การทำนายการเกิดภาวะเลือดออกในสมองจากการใช้ยาละลายลิ่มเลือดในผู้ป่วยโรคหลอดเลือดสองตีบและอุดตันโดยใช้แบบจำลองโครงข่ายประสาทเทียมหลายแบบ ภาควิชากายวิภาคศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยมหิดล และภาควิชาอายุรศาสตร์ คณะแพทยศาสตร์ มหาวิทยาลัยธรรมศาสตร์. Neurological Research 34 (2012) p. 120-128.

15) เสกสันติ จันทะมงคล สมจิตร อาจอินทร์ งามนิจ อาจอินทร์ และวัชชพล เดชขันธ์. 2555. ระบบช่วยเหลืออัจฉริยะเพื่อการวางแผนการรักษาโรคเรื้อรังโดยใช้เหมืองข้อมูล. ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น. NCIT2012, 26-27 Aprilm Cha-am, Thailand, p 117-128.

16) วีระยุทธ มายุศิริ จารี ทองคำและวิทินี สุขมาก. 2556. การพัฒนาแบบจำลองเพื่อการพยากรณ์การรักษาซ้ำของผู้ป่วยโรคจิตเภทโดยเทคนิคเหมืองข้อมูล. วารสารวิทยาศาสตร์และเทคโนโลยี. การประชุมวิชาการ มหาสารคามวิจัย ครั้งที่ 10. หน้า 144-153.

17) เชาวนันท์ โสโท พุธษดี ศิริแสดงตระกูล และวรชัย ตั้งวรพงศ์ชัย. 2556. แบบจำลองการทำนายผลการรักษาผู้ป่วยมะเร็งปากมดลูกด้วยโครงข่ายประสาทเทียม. ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น. วารสารวิจัย มข. (บศ.) 13 (1) : มค. – มี.ค. 2556. หน้า 39-50.

18) Priya, R. and Aruna, P. SVM and neural network based diagnosis of diabetic retinopathy. International Journal of Computer Applications (0975-8887), Volume 41-No. 1, March 2012 : 6-12.

เอกสารอ้างอิง (ต่อ)

19) Sa-ngasoongsong, A. and Chongwatpol, J. An analysis of diabetes risk factors using data mining approach. Oklahoma State University, Stillwater, OK 74078, USA, Paper PH10-2012 : 1-11.

20) Al-Rofiyee, A., Al-Nowiser, M., Al-Mufadi, N., AL-Hagery, M. A. Using prediction methods in data mining for diabetes diagnosis. Qassim University, College of Computer, Department of IT, Kingdom of Saudi Arabia, 2013, Poster : 1-2.

21) Cela, E. and Frasheri, N. Data mining techniques and tools used in healthcare databases. The 1st International Conference on "Research and Education-Challenges Towards the Future" (ICRAE2013), 24-25, May 2013 : 1-11.

22) Folorunsho, O. Comparative study of different data mining techniques performance in knowledge discovery from medical database. International Journal of Advanced Research in Computer Science and Software Engineering. Volume 3, Issue 3, March 2013 : 11-15.

23) Rahman, R. M. and Afroz, F. Comparison of various classification techniques using differential data mining tools for diabetes diagnosis. Journal of Software Engineering and Applications, 2013, 6, 85-97.

24) Ananthapadmanaban, K. R. and Parthiban, G. Prediction of changes-diabetic retinopathy using data mining classification techniques. Indian Journal of Science and Technology, Vol.7(10), 1498-1503, October, 2014.

25) Giri, T. N. and Todamal, S. R. Data mining approach for diagnosing type 2 diabetes. International Journal of Science, Engineering and Technology. Vol. 2, Issue 8, November-December, 2014, 191-194.

26) Kumar, S. and Gayathri, P. Analysis of adult-onset diabetes using data mining classification algorithms. International Journal of Modern Computer Science (IJMCS). Volume No.2, Issue No.3, Jun, 2014, Conference proceeding, 37-42.

27) Kumar, V. P. and Velide, L. A data mining approach for prediction and treatment of diabetes disease. International Journal of Science Inventions Today, Volume 3, Issue 1, January-February 2014.

เอกสารอ้างอิง (ต่อ)

28) Sanakal, R. and Jayakumari. S. T. Prognosis of diabetes using data mining approach-Fuzzy C Means clustering and support vector machine. International Journal of Computer Trends and Technology (IJCTT). Volume 11(2), May 2014 : 94-98.

29) Agrawal, P. and Dewangan, A. K. A Brief survey on the techniques used for the diagnosis of diabetes-mellitus. International Research Journal of Engineering and Technology (IRJET). Volume 02 Issue 03, June, 2015 : 1039-1043.

30) Iyer, A., Jeyalatha, S. and Sumbaly, R. Diagnosis of Diabetes using classification mining techniques. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015 : 1-14.

31) Devi, M. R. Analysis of various data mining techniques to predict diabetes Mellitus. International Journal of Applied Engineering Research ISSN 0973-4562. Volume 11, Number 1 (2016) : 727-730.

32) พยูน พาณิชย์กุล. 2548. การพัฒนาระบบดาต้าไมน์นิ่งโดยใช้ Decision Tree. โครงการพัฒนาระบบงานปริญญาวิทยาศาสตร์มหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ แขนงวิทยาการสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง.

33) Kumar, V. and Wu, X. 2009. The Top Ten Algorithms in Data Mining. University of Minnesota Department of Computer Science and Engineering, Minneapolis, Minnesota : CRC Press.

34) รุจิรา ธรรมสมบัติ. 2554. ระบบสนับสนุนการตัดสินใจในการเลือกใช้แพคเกจอินเทอร์เน็ตมือถือโดยใช้ต้นไม้ตัดสินใจ. สาขาคอมพิวเตอร์ธุรกิจ คณะบริหารธุรกิจ วิทยาลัยราชพฤกษ์.

35) Berson, A. and Stephen, J. S, 1997. Data Warehousing, Data Mining, and OLAP.

36) กัลยา วานิชย์บัญชา. 2552. การวิเคราะห์ข้อมูลหลายตัวแปร. กรุงเทพฯ : บริษัทธรรมสาร จำกัด.

37) สายชล สินสมบูรณ์ทอง. 2558. การทำเหมืองข้อมูล. กรุงเทพฯ : จามจุรีโปรดักส์ จำกัด.

เอกสารอ้างอิง (ต่อ)

38) ณัฐวุฒิ ศิริกุลรุ่งโรจน์ และคณะ. 2556. การเปรียบเทียบประสิทธิภาพการจำแนกประเภทโดยใช้วิธีความใกล้เคียงกันมากที่สุด วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ วิธีโครงข่ายประสาทเทียมและวิธีซัพพอร์ตเวกเตอร์แมชชีน. ปัญหาพิเศษ ปริญญาวิทยาศาสตรบัณฑิต สาขาวิชาสถิติ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง.