

การพัฒนาและเปรียบเทียบแบบจำลองเหมืองข้อมูลกับสมการถดถอยเชิงเส้น พหุคูณสำหรับพยากรณ์ปริมาณโรคอ้วนแบบอัตโนมัติ

ประภาพรรณ ยาลังกา¹ และ เทวิน ธนวงษ์¹

¹ภาควิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยนครสวรรค์ พิษณุโลก

Emails: prapapany56@email.nu.ac.th, tawint@nu.ac.th

บทคัดย่อ

วิทยานิพนธ์นี้ได้ทำการพัฒนาและเปรียบเทียบแบบจำลองเหมืองข้อมูลกับสมการถดถอยเชิงเส้นพหุคูณ สำหรับพยากรณ์ปริมาณโรคอ้วนแบบอัตโนมัติ โดยใช้เทคนิคต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 (หรือ J48) อัลกอริทึม J48graft และอัลกอริทึม ID3 และเทคนิคโครงข่ายประสาทเทียมแบบมีลติเลเยอร์เพอร์เซ็ปตรอน ในการสร้างโมเดล แบ่งข้อมูลด้วยวิธีการแบ่งข้อมูลแบบสุ่มด้วยการแบ่งร้อยละ (Percentage Split) ในอัตราส่วน 80:20 และ 34:66 และการตรวจสอบไขว้แบบ 5-Fold Cross Validation และ 10-Fold Cross Validation โดยแบ่งข้อมูลชุดเรียนรู้และชุดทดสอบในอัตราส่วน 70:30 ได้โมเดลพยากรณ์มาจำนวน 24 โมเดล แล้วเปรียบเทียบค่าความถูกต้องของแต่ละโมเดลกับค่า Adjusted R Square ที่ได้จากการทำ Regression analysis ผลการศึกษาพบว่า เมื่อเปรียบเทียบแล้วอัลกอริทึม ID3 ด้วยวิธีการแบ่งข้อมูลแบบตรวจสอบไขว้ ชนิด 10-Fold มีค่าความถูกต้องในการจำแนกข้อมูลสูงสุด เท่ากับ 100% จึงเลือกใช้โมเดลที่ได้จากอัลกอริทึม ID3 ด้วยวิธีการแบ่งข้อมูลแบบตรวจสอบไขว้ 10-Fold มาใช้ในการพัฒนาระบบพยากรณ์ปริมาณโรคอ้วนแบบอัตโนมัติต่อไป

ABSTRACT

This research had developed and compared the data mining model with the multiple regression for automatic forecasting the obesity. The decision tree which consisted of algorithm C45 (or J48), algorithm J48 graft and algorithm ID3, and the multilayer

perceptron of the artificial neural network were used to create the model. The data were divided by the random data method which was percentage split at 80:20 and 34:66 ratio, and the 5-Fold and 10-Fold cross validation, which were categorized in learning data set and testing data set at 70:30 ratio. This created 24 forecasting models. Then each model was compared the accuracy with Adjusted R Square, which was from regression analysis. The research result shown that, after comparing the algorithm ID3 by cross validation, the accuracy level of highest classified data of 10-Fold was 100 percent. Therefore, the model of the algorithm ID3 by 10-fold cross validation was chosen to use for development in the further system of automatic forecasting for obesity.

คำสำคัญ— ต้นไม้ตัดสินใจ; โครงข่ายประสาทเทียม; โรคอ้วน; สมการถดถอยเชิงเส้นพหุคูณ

1. บทนำ

ปัจจุบันประเทศไทยกำลังเผชิญปัญหาสุขภาพโภชนาการในเด็กไทย พบว่ามีภาวะโภชนาการเกินและเป็นโรคอ้วนมากขึ้นพบว่าโรคอ้วนและโรคที่เกี่ยวข้องกับโรคอ้วน เช่น เบาหวาน ไขมันในเลือดสูง หลอดเลือดแข็ง เป็นต้น ซึ่งเป็นโรคที่เริ่มต้นจากวัยเด็กแล้วค่อยมีอาการแสดง เห็นชัดเมื่อเข้าสู่วัยผู้ใหญ่และอุบัติการณ์โรคอ้วนในเด็กก็เพิ่มขึ้นทั่วโลกเช่นเดียวกัน สภาพสังคมที่เปลี่ยนไป

ทำให้เด็กมีพฤติกรรมอาการที่ไม่พึงประสงค์มากยิ่งขึ้น ไม่ค่อยออกกำลังกาย ซึ่งเป็นปัจจัยเสี่ยงต่อการเกิดโรคเรื้อรังไม่ติดต่อหลายชนิดเมื่อเติบโตขึ้น เป็นภาระในการดูแลรักษา ทั้งในระดับครอบครัว และประเทศ รัฐจะต้องลงทุนด้านการบริการทางการแพทย์มากขึ้น การสำรวจพบว่าเด็กไทยมีพฤติกรรมอาการหวานเพิ่มขึ้น ซึ่งเป็นแหล่งที่มาของน้ำตาลได้แก่ เครื่องดื่มประเภทน้ำหวาน น้ำอัดลม ขนมกรุบกรอบที่ส่งผลต่อภาวะโภชนาการเกิน ผู้ศึกษาเห็นความสำคัญของปัจจัยด้านต่างๆ ที่มีผลต่อปริมาณการเกิดโรคอ้วนในเด็กนักเรียนชั้นประถมศึกษา ผู้ศึกษามีความสนใจที่จะประยุกต์ใช้เทคนิคเหมืองข้อมูลในการพัฒนาและเปรียบเทียบแบบจำลองเหมืองข้อมูลกับสมการถดถอยเชิงเส้นพหุคูณสำหรับพยากรณ์ปริมาณโรคอ้วนแบบอัตโนมัติ ซึ่งเป็นการช่วยให้ทราบถึงแนวโน้มของปริมาณโรคอ้วนในนักเรียนชั้นประถมศึกษาในจังหวัดพิษณุโลก เพื่อเป็นประโยชน์แก่โรงเรียนหรือกรมอนามัยในการช่วยเหลือปรับเปลี่ยนพฤติกรรมเสี่ยงต่างๆ และเป็นแนวทางในการออกแบบกิจกรรมส่งเสริมสุขภาพเด็กนักเรียนระดับประถมศึกษาให้นักเรียนมีสุขภาพที่ดีและแข็งแรงขึ้น

2. วัตถุประสงค์ของการศึกษา

- 1). เพื่อเปรียบเทียบแบบจำลองการวิเคราะห์สมการถดถอยเชิงเส้นพหุคูณกับเทคนิคเหมืองข้อมูล สำหรับพยากรณ์ปริมาณโรคอ้วนแบบอัตโนมัติ
- 2). เพื่อสร้างเว็บแอปพลิเคชัน สำหรับพยากรณ์ปริมาณโรคอ้วนแบบอัตโนมัติ โดยผ่าน Responsive Web Application
- 3). เพื่อฝึกทักษะการเขียนโปรแกรมและการประยุกต์ใช้ทาง Data Mining

3. เอกสารและงานวิจัยที่เกี่ยวข้อง

3.1 สมการถดถอยเชิงเส้นพหุคูณ

สมการถดถอยเชิงเส้นพหุคูณ [6] คือการศึกษาความสัมพันธ์ระหว่างตัวแปรอิสระหรือตัวแปรต้นที่ทำหน้าที่พยากรณ์ตั้งแต่ 2 ตัวขึ้นไป กับตัวแปรตาม 1 ตัว ในการวิเคราะห์การถดถอยพหุคูณนั้นจะต้องหาค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณ เพื่อให้ทราบถึงความสัมพันธ์ระหว่างตัวแปรอิสระหรือตัวแปรต้นกับตัวแปรตามว่ามีความสัมพันธ์กันเช่นใด

3.2 การทำเหมืองข้อมูล

การทำเหมืองข้อมูล [7] เป็นกระบวนการจัดการกับข้อมูลจำนวนมากเพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อน อยู่ในชุดข้อมูลนั้นในปัจจุบันการทำเหมืองข้อมูลได้ถูกนำไปประยุกต์ใช้ในงานหลายประเภท ทั้งในด้านธุรกิจที่ช่วยในการตัดสินใจของผู้บริหาร และในด้านวิทยาศาสตร์และการแพทย์ รวมทั้งในด้านเศรษฐกิจและสังคม

3.2.1 เทคนิคการจัดกลุ่ม (Clustering)

เทคนิคการจัดกลุ่ม [5] จะแตกต่างจากการแบ่งประเภทข้อมูล (Classification) โดยจะแบ่งกลุ่มข้อมูลจากความคล้าย โดยไม่มีการกำหนดคลาสประเภทข้อมูลไว้ก่อนหรือไม่ทราบจำนวนกลุ่มล่วงหน้า เป็นการเรียนรู้แบบไม่มีผู้สอน และเป็นการแบ่งกลุ่มที่ต้องใช้ข้อมูลของทุกกลุ่มในการหาจุดเด่นของแต่ละกลุ่มออกมาให้เห็นอย่างชัดเจน เพื่อได้มาซึ่งลักษณะเฉพาะของกลุ่มนั้นๆ

3.2.2 เทคนิคต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจ [9] เป็นการนำข้อมูลมาสร้างแบบจำลองการพยากรณ์ในรูปแบบโครงสร้างต้นไม้ตัดสินใจ ซึ่งเป็นลักษณะการทำงานแบบการเรียนรู้แบบมีผู้สอน มีลักษณะคล้ายต้นไม้จริงกลับหัวที่มีโหนดรากอยู่ด้านบนสุดและโหนดใบอยู่ล่างสุดของต้นไม้ ภายในต้นไม้จะประกอบไปด้วยโหนด (node) ซึ่งแต่ละโหนดจะมีคุณลักษณะเป็นตัวทดสอบ กิ่งของต้นไม้ (branch) แสดงถึงค่าที่เป็นไปได้ของคุณลักษณะที่ถูกเลือกทดสอบ และใบ (leaf) ซึ่งเป็นสิ่งที่อยู่ล่างสุดของต้นไม้ตัดสินใจแสดงถึงกลุ่มของข้อมูล (class) หรือนั่นก็คือผลลัพธ์ที่ได้จากการทำนาย โหนดที่อยู่บนสุดของต้นไม้เรียกว่าโหนดราก (root node)

3.2.3 เทคนิคโครงข่ายประสาทเทียม

เทคนิคโครงข่ายประสาทเทียม (Neural Networks) [4] คือโมเดลทางคณิตศาสตร์ ใช้สำหรับประมวลผลสารสนเทศด้วยการคำนวณในแบบคอนเนกชันนิสต์ (connectionist) เพื่อจำลองการทำงานของเครือข่ายประสาทในสมองมนุษย์ เป็นเทคโนโลยีที่มีที่มาจากงานวิจัยด้านปัญญาประดิษฐ์ เพื่อใช้ในการคำนวณค่าฟังก์ชันจากกลุ่มข้อมูล วิธีการของนิวรอลเน็ตเป็นวิธีการที่ให้

เครื่องเรียนรู้จากตัวอย่างต้นแบบ แล้วฝึก (train) ให้ระบบได้รู้จักที่จะคิดแก้ปัญหาที่กว้างขึ้นได้ ในโครงสร้างของนิวรอลเน็ตจะประกอบด้วยโหนด (node) สำหรับ Input – Output และการประมวลผลกระจายอยู่ในโครงสร้างเป็นชั้นๆ ได้แก่ input layer output layer และ hidden layers การประมวลผลจะอาศัยการส่งการทำงานผ่านโหนดต่าง ๆ ใน layer เหล่านี้

3.3 โรคอ้วน (obesity)

โรคอ้วน [8] หมายถึง ภาวะที่ร่างกายมีการสะสมไขมันในส่วนต่างๆของร่างกายเกินปกติ จนเป็นปัจจัยเสี่ยง หรือเป็นสาเหตุให้เกิดโรคต่างๆที่ส่งผลถึงสุขภาพ จนอาจเป็นสาเหตุให้เสียชีวิตได้ โดยเมื่อมีค่าดัชนีมวลกาย (BMI/บีเอ็มไอ) ตั้งแต่ค่า 25 ขึ้นไป เรียกว่า น้ำหนักตัวเกิน แต่ถ้ามีค่าดัชนีมวลกายตั้งแต่ 30 ขึ้นไป เรียกว่า เป็นโรคอ้วน

3.4 Weka 3.6.12

Weka สำหรับทำ Data Mining [10] เป็นซอฟต์แวร์ด้านการทำเหมืองข้อมูลที่ได้รับการยอมรับและแพร่หลายในต่างประเทศ รวมทั้งประเทศไทยด้วย ซึ่งเห็นได้จากงานวิจัยหลายเรื่องที่นำมาโปรแกรมเวก้ามาใช้ในการพัฒนารูปแบบจำลองต่างๆ สาเหตุอีกประการหนึ่งที่ทำให้ได้รับความนิยมก็เป็นซอฟต์แวร์เปิด ที่อยู่ภายใต้การควบคุมของ GPL License ทำให้มีผู้ใช้หลากหลายทั่วไปทุกระดับ เป็นเครื่องมือที่ใช้ทำงานในด้านการทำดาต้าไมนิ่งที่รวบรวมแนวคิดอัลกอริทึมมากมาย

3.5 งานวิจัยที่เกี่ยวข้อง

กรวิภา ภูณพมา (2555) ได้คิดค้นระบบวิเคราะห์ความเสี่ยงของการเกิดโรคอ้วนลงพุง [1] โดยประยุกต์ใช้เทคนิคการทำเหมืองข้อมูล (Data Mining) ซึ่งเทคนิคที่นำมาใช้คือ ต้นไม้ตัดสินใจ (Decision Tree) ผลการพัฒนาระบบพบว่า ระบบสามารถวิเคราะห์ความเสี่ยงของการเกิดโรคอ้วนลงพุงและแนะนำอาหารสำหรับผู้ป่วยได้อย่างมีประสิทธิภาพ ส่วนการทดสอบความถูกต้องของโมเดลโดยใช้โปรแกรม Weka 3.7 พบว่าโมเดลของการวิเคราะห์ความเสี่ยงของการเกิดโรคอ้วนลงพุงมีค่าความถูกต้องเท่ากับ 97.36% และโมเดลของการแนะนำอาหารสำหรับผู้ป่วยมีค่าความถูกต้องเท่ากับ 87.92% ดังนั้นระบบที่พัฒนาขึ้นนี้มีผลประเมินค่าความพึงพอใจในระดับดี และสามารถ

นำไปประยุกต์ใช้งานได้ นอกจากนี้ยังพบว่างานวิจัยของวงกต ศรีอุไร (2557) ได้ศึกษาการจำแนกผู้ป่วยโรคอ้วนลงพุงโดยใช้วิธีการคัดเลือกคุณลักษณะและโครงข่ายประสาทเทียม[3] โดยสร้างโมเดลในการจำแนกข้อมูลผู้ป่วยโรคอ้วนลงพุงโดยใช้วิธีการคัดเลือกคุณลักษณะและโครงข่ายประสาทเทียมแบบมัลติเลเยอร์เพอร์เซ็ปตรอน สำหรับข้อมูลที่ใช้ได้มาจากการเก็บข้อมูลโดยใช้แบบสอบถามจากบุคลากรในโรงพยาบาลค่ายสรรพสิทธิประสงค์ตั้งแต่ปี พ.ศ. 2550-2554 ผลการวิจัยพบว่าการใช้วิธีการคัดเลือกคุณลักษณะด้วย Correlation-based Feature Selection ร่วมกับโครงข่ายประสาทเทียมแบบมัลติเลเยอร์เพอร์เซ็ปตรอนในการสร้างโมเดลการจำแนกข้อมูลจะให้ค่าความถูกต้องในการจำแนกข้อมูลสูงที่สุดซึ่งมีค่าเท่ากับ 92.56% สำหรับงานวิจัยของเกษณี สุขพิมาย (2555) ได้ศึกษาปัจจัยที่มีความสัมพันธ์กับพฤติกรรมการบริโภคอาหารและการออกกำลังกายของกลุ่มเสี่ยงโรคอ้วนลงพุง [2] โดยศึกษากลุ่มเสี่ยงโรคอ้วนลงพุงอายุ 35 – 60 ปี ผลการศึกษาพบว่า กลุ่มเสี่ยงโรคอ้วนลงพุงร้อยละ 56.2 มีพฤติกรรมการบริโภคอาหารในระดับปรับปรุงร้อยละ 70.2 มีพฤติกรรมการออกกำลังกายในระดับปรับปรุงปัจจัยที่มีความสัมพันธ์กับพฤติกรรมการบริโภคอาหารของกลุ่มเสี่ยงโรคอ้วนลงพุงประกอบด้วย การรับรู้ความสามารถตนเอง ($p < 0.001$) สำหรับปัจจัยเสริมที่สัมพันธ์กับพฤติกรรมการออกกำลังกาย ได้แก่ การรับรู้ความสามารถตนเอง ($p < 0.001$) การสนับสนุนทางสังคม จากสมาชิกในครอบครัว ($p < 0.001$) เพื่อน/เพื่อนบ้าน ($p < 0.001$) และจากเจ้าหน้าที่สาธารณสุข ($p < 0.002$) ดังนั้น การจัดโครงการกิจกรรมเพื่อส่งเสริมพฤติกรรม ควรเน้นการสร้างการรับรู้ความสามารถตนเองในการปฏิบัติ รวมถึงเพิ่มการสนับสนุนจากสมาชิกในครอบครัว และเจ้าหน้าที่สาธารณสุข

4. วิธีการดำเนินการศึกษาวิจัย

4.1 ศึกษาและเก็บรวบรวมข้อมูล

ผู้วิจัยได้ทำการศึกษาและรวบรวมข้อมูลต่างๆที่เกี่ยวข้องกับงานวิจัย โดยเก็บข้อมูลของนักเรียนระดับประถมศึกษาปีที่ 1-6 ในจังหวัดพิษณุโลก โดยใช้แบบสอบถาม ทั้งหมด 965 ชุดข้อมูล

4.2 คัดเลือกและจัดเตรียมข้อมูล

1) การคัดเลือกข้อมูล จากข้อมูลที่ได้รับรวบรวมมาจำนวน 965 ชุดข้อมูล ผู้วิจัยนำมาทำการแบ่งกลุ่มข้อมูล (Clustering) โดยจะแบ่งกลุ่มข้อมูลจากความคล้ายกันอยู่กลุ่มเดียวกัน ด้วยวิธีการจัดกลุ่มแบบ Simple K Means โดยใช้ $K = 4$ (4 กลุ่ม) เมื่อทำการแบ่งกลุ่มข้อมูลแล้วจึงได้ข้อมูลทั้งหมด 772 ชุดข้อมูล แบ่งข้อมูลออกเป็นสองชุด เพื่อใช้ในการสร้างโมเดล(Train) และทดสอบโมเดล(Test) ในอัตราส่วน 70:30

2) การจัดเตรียมข้อมูล โดยจะทำการแปลงข้อมูลจากไฟล์นามสกุล .csv ให้อยู่ในรูปแบบของไฟล์นามสกุล .arff

4.3 สร้างโมเดลเพื่อพยากรณ์

โดยใช้เทคนิคต้นไม้ตัดสินใจ และเทคนิคโครงข่ายประสาทเทียม นั้น ผู้วิจัยเลือกใช้โปรแกรมเวก้า (WEKA) ในการสร้างโมเดลพยากรณ์ โดยเทคนิคต้นไม้ตัดสินใจนั้นได้ใช้อัลกอริทึม C4.5 (หรือ J48), J48graft และอัลกอริทึม ID3 ส่วนเทคนิคโครงข่ายประสาทเทียมนั้น ใช้ Multilayer Perceptron โดย hidden Layers เป็น 2, 3 และ 5 ตามลำดับ

ผู้วิจัยได้เลือกวิธีสร้างโมเดลและทดสอบโมเดล 2 วิธี คือ วิธีการแบ่งข้อมูลแบบสุ่มด้วยการแบ่งร้อยละ และวิธีการตรวจสอบไขว้ แบบ 5-Fold และ 10-Fold ทำการแบ่งข้อมูลออกเป็น 2 ชุด ในการสร้างโมเดล(Train) และทดสอบโมเดล (Test) ในอัตราส่วน 80:20 และ 34:66

ส่วนของ Multiple regression โดยข้อมูลทั้งหมดต้องเป็นตัวเลข แปลงข้อมูลจากข้อมูลเชิงคุณภาพเป็นข้อมูลเชิงปริมาณ ได้มาจากการทำ Dummy Variables และทำการ Regression จากนั้นดูค่า Adjusted R Square ที่ได้เลือกข้อมูลที่มีค่า Adjusted R Square ที่มากที่สุด มาใช้ในการเปรียบเทียบโมเดล

4.4 ทดสอบความถูกต้องของโมเดลเพื่อพยากรณ์

ผู้วิจัยได้ใช้เทคนิคต้นไม้ตัดสินใจด้วยอัลกอริทึม J48, J48graft และ ID3 และเทคนิคโครงข่ายประสาทเทียม ด้วยวิธีการแบ่งข้อมูลแบบสุ่มด้วยการแบ่งร้อยละ ในอัตราส่วน 80:20 และ 34:66 และวิธีการตรวจสอบไขว้แบบ 5-Fold และ 10-Fold โดยแบ่งข้อมูลเป็น 2 ชุด ในอัตราส่วน 70:30 เมื่อทำการสร้างโมเดลการพยากรณ์เสร็จเรียบร้อยแล้ว จะต้องทำการทดสอบความถูกต้องในการทำนายของโมเดลการพยากรณ์ โดยการนำข้อมูลชุด

ทดสอบ (Test) ที่แบ่งออกไว้ มาใช้ในการทดสอบโมเดลการพยากรณ์

4.5 วิเคราะห์และออกแบบระบบ

จากผลการวิเคราะห์โมเดลที่สร้างด้วยการแบ่งข้อมูลแบบการตรวจสอบไขว้ (10-Fold Cross Validation) ซึ่งเป็นวิธีที่มีประสิทธิภาพ มีค่าประสิทธิภาพความถูกต้อง เท่ากับ 100% ทำให้ทราบแอททริบิวต์ที่จะนำมาใช้ในการพยากรณ์โดยมีจำนวนกฎที่ใช้ในการพยากรณ์ 283 กฎ และได้ออกแบบระบบดังนี้

ส่วนของผู้ใช้

1. สามารถรอกข้อมูลที่ใช้ต้องการพยากรณ์ปริมาณโรคอ้วนของเด็กระดับประถมศึกษาได้
2. สามารถทราบได้ว่า ปริมาณโรคอ้วนในเด็กประถมศึกษาของแต่ละโรงเรียนเป็นอย่างไร
3. สามารถพยากรณ์ได้ว่า มีจำนวนนักเรียนที่อ้วนกี่คนของทั้งหมดในโรงเรียน
4. สามารถคำนวณเปอร์เซ็นต์ความคลาดเคลื่อนในการพยากรณ์ได้

4.6 พัฒนาระบบ

ผู้วิจัยได้ทำการพัฒนาเว็บแอปพลิเคชันด้วยภาษา PHP, HTML, CSS และ JavaScript และนำกฎที่ได้จากการสร้างและทดสอบโมเดล ด้วยเทคนิคต้นไม้ตัดสินใจด้วยอัลกอริทึม ID3 จำนวน 283 กฎ พัฒนาเป็นเงื่อนไขในการวิเคราะห์คลาสผลลัพธ์

5. ผลการศึกษา

5.1 ผลการสร้างโมเดล

โดยอัลกอริทึม J48 และ J48graft ด้วยวิธีการแบ่งข้อมูลแบบสุ่มด้วยการแบ่งร้อยละ 20 และ 66 มีค่าความถูกต้องอยู่ที่ 79.13% 79.39% ตามลำดับ และอัลกอริทึม ID3 มีค่าความถูกต้องของโมเดลอยู่ที่ 61.16% และ 70.61% ตามลำดับ และอัลกอริทึม J48 และ J48graft ด้วยวิธีการตรวจสอบไขว้ ชนิด 5-Fold และ 10-Fold มีค่าความถูกต้องอยู่ที่ 86.64% และ 86.64% ตามลำดับ โดยอัลกอริทึม ID3 มีค่าความถูกต้องของทั้งสองโมเดลอยู่ที่ 100.00%

ส่วนเทคนิคโครงข่ายประสาทโดยโมเดลการพยากรณ์ที่ใช้ Hidden node2 Hidden node3 และ Hidden node5 ด้วยวิธีการแบ่งข้อมูลแบบสุ่มด้วยการแบ่งร้อยละ 20 มีค่าความถูกต้องอยู่ที่ 75.57% 75.40% และ 92.88% ตามลำดับ ส่วนการแบ่งร้อยละ 66 มีค่าความถูกต้องอยู่ที่ 81.68% 83.21% และ 96.95% ตามลำดับ และด้วยวิธีการตรวจสอบไขว้ ชนิด 5-Fold และ 10-Fold มีค่าความถูกต้องเท่ากันอยู่ที่ 86.64% 88.36% และ 99.14% ตามลำดับ

5.2 วิเคราะห์ เปรียบเทียบและอภิปรายผล

ตาราง 1. ค่าประสิทธิภาพความถูกต้องที่ได้จากการทดสอบโมเดลการพยากรณ์ต้นไม้ตัดสินใจ

รูปแบบของแบบจำลองต้นไม้ตัดสินใจ			ค่าความถูกต้อง (%)	ค่าความแม่นยำ (%)	ค่าความระลึก (%)
Decision Tree	J48	Cross-validation (Fold=5)	86.64	82.00	93.60
		Cross-validation (Fold=10)	86.64	82.00	93.60
		Percentage split (20)	79.13	73.00	89.20
		Percentage split (66)	79.39	71.60	85.00
	J48graft	Cross-validation (Fold=5)	86.64	82.00	93.60
		Cross-validation (Fold=10)	86.64	82.00	93.60
		Percentage split (20)	79.13	73.00	89.20
		Percentage split (66)	79.39	71.60	85.00
	ID3	Cross-validation (Fold=5)	100.00	100.00	100.00
		Cross-validation (Fold=10)	100.00	100.00	100.00
		Percentage split (20)	61.16	70.40	66.50
		Percentage split (66)	70.61	74.40	83.10

สรุปได้ว่าโมเดลที่สร้างด้วยอัลกอริทึม ID3 ด้วยวิธีการตรวจสอบไขว้ ชนิด 5-Fold และ 10-Fold มีประสิทธิภาพสูงสุดโดยมีค่าความถูกต้องเท่ากับ 100.00%

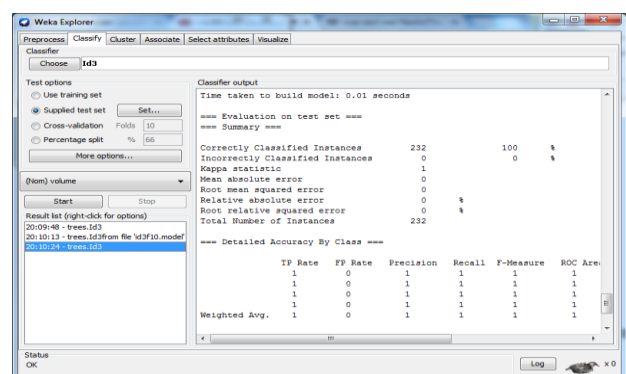
ตาราง 2. ค่าประสิทธิภาพความถูกต้องที่ได้จากการทดสอบโมเดลการพยากรณ์โครงข่ายประสาทเทียม

รูปแบบของแบบจำลองโครงข่ายประสาทเทียม			ค่าความถูกต้อง (%)	ค่าความแม่นยำ (%)	ค่าความระลึก (%)
ANN	HN2	Cross-validation (Fold=5)	86.64	100.00	94.30
		Cross-validation (Fold=10)	86.64	100.00	94.30
		Percentage split (20)	75.57	95.40	100
		Percentage split (66)	81.68	100.00	92.40
	HN3	Cross-validation (Fold=5)	88.36	100.00	100.00
		Cross-validation (Fold=10)	88.36	100.00	100.00
		Percentage split (20)	75.40	96.00	100.00
		Percentage split (66)	83.21	97.10	100.00
	HN5	Cross-validation (Fold=5)	99.14	97.20	100.00
		Cross-validation (Fold=10)	99.14	97.20	100.00
		Percentage split (20)	92.88	95.40	100.00
		Percentage split (66)	96.95	97.10	100.00

และเทคนิคโครงข่ายประสาทเทียมที่ใช้ Hidden node5 ด้วยวิธีการตรวจสอบไขว้แบบชนิด 5-Fold และ 10-Fold มีประสิทธิภาพสูงสุด โดยมีค่าความถูกต้องเท่ากับ 99.14 %

ตาราง 3. ค่าประสิทธิภาพต่างๆของการทำ Multiple regression

	F	Sig.	Adjusted R Square (%)
Multiple regression	206.94	0.0005	76.22



รูปที่ 1. โมเดลโดยใช้อัลกอริทึม ID3 แบ่งข้อมูลแบบ 10-Fold Cross Validation

```

45 <?php
46 $a = $_POST['a'];
47 $a = $_POST['a'];
48 $l = $_POST['l'];
49 $w = $_POST['w'];
50 $h = $_POST['h'];
51 $num = $_POST['num'];
52 $n = $_POST['n'];
53 $di = $_POST['di'];
54 $c = $_POST['c'];
55 $e = $_POST['e'];
56 $m = $_POST['m'];
57 $fa = $_POST['fa'];
58 $so = $_POST['so'];
59 $ca = $_POST['ca'];
60 $f = $_POST['f'];
61 $la = $_POST['la'];
62 $st = $_POST['st'];
63 $cr = $_POST['cr'];
64 $p = $_POST['p'];
65
66
67 if ($a == 'a1' && $e == 'e4' && $cr == 'cr4' && $s == 's2') {
68     $re = "หน้าจอนักเรียนและคุณครู";
69     $num = round($num/23/244);
70 }
71 else if ($a == 'a1' && $e == 'e4' && $cr == 'cr4' && $s == 's1') {
72     $re = "หน้าจอนักเรียนและคุณครู";
73     $num = round($num/15/174);
74 }
75 else if ($a == 'a1' && $e == 'e4' && $cr == 'cr5') {
76     $re = "หน้าจอนักเรียนและคุณครู";
77     $num = round($num/23/244);
78 }
79 else if ($a == 'a1' && $e == 'e4' && $cr == 'cr3' && $l == '15') {
80     $re = "หน้าจอนักเรียนและคุณครู";
81     $num = round($num/23/244);
82 }
83 else if ($a == 'a1' && $e == 'e4' && $cr == 'cr3' && $l == '15') {
84     $re = "หน้าจอนักเรียนและคุณครู";
85     $num = round($num/23/244);
86 }
    
```

รูปที่ 2. หน้าจอการเขียนกฎจากโมเดล ID3 ด้วยภาษา PHP

5. สรุปผลการดำเนินการ

จากการศึกษาพบว่า โมเดลอัลกอริทึม ID3 ด้วยวิธีการตรวจสอบไขว้ แบบ 5-Fold และวิธีการตรวจสอบไขว้ แบบ -10Fold มีประสิทธิภาพสูงสุด โดยมีค่าความถูกต้อง เท่ากับ 100.00% จึงได้นำเอาข้อมูลที่ได้จากการพยากรณ์ของอัลกอริทึม ID3 มาใช้ประกอบการทำรายงานและแสดงผลต่อไป ผลการทดลอง พบว่า การพยากรณ์ปริมาณโรคอ้วนแบบอัตโนมัติ ของเด็ก ประถมศึกษาในจังหวัดพิษณุโลก ด้วยเทคนิคเหมืองข้อมูล ที่พัฒนาขึ้นนั้น ได้บรรลุวัตถุประสงค์ ซึ่งเป็นรูปแบบของเว็บแอปพลิเคชันแบบ Responsive ที่สามารถเข้าถึงได้ง่าย และดูข้อมูลที่ได้จากการพยากรณ์ได้อย่างสะดวก

รูปที่ 3. หน้าจอการวิเคราะห์ปริมาณโรคอ้วน

รูปที่ 4. หน้าจอผลลัพธ์การพยากรณ์และวิเคราะห์ ความคลาดเคลื่อนของการพยากรณ์ปริมาณโรคอ้วน

เอกสารอ้างอิง

- [1] กรวิกา ภูวนพา และคณะ, “ระบบวิเคราะห์ความเสี่ยงของการเกิดโรคอ้วนลงพุงและแนะนำอาหารสำหรับผู้ป่วย”, สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ.
- [2] เกษณี สุขพิมาย และคณะ, “ปัจจัยที่มีความสัมพันธ์กับพฤติกรรมบริโภคอาหารและการออกกำลังกายของกลุ่มเสี่ยงโรคอ้วนลงพุง” วารสารสุขศึกษา ปีที่ 35 เล่มที่ 122.
- [3] วงศ์ ศรีอุไร, “การจำแนกผู้ป่วยโรคอ้วนลงพุงโดยใช้วิธีการคัดเลือกคุณลักษณะและโครงข่ายประสาทเทียม” วารสารวิทยาศาสตร์ มศว ปีที่ 30 ฉบับที่ 1.
- [4] วิทยา พรพิจรพงศ์, โครงข่ายประสาทเทียม (Artificial Neural Networks - ANN) สืบค้นเมื่อ 11 เมษายน 2559, จาก <https://www.gotoknow.org/posts/163433>
- [5] วิภาวรรณ บัวทอง, เทคนิคการจัดกลุ่ม (Clustering) สืบค้นเมื่อ 30 มีนาคม 2559, จาก <https://wipawanblog.files.wordpress.com/2014/06/chapter-8-clustering-k-means.pdf>
- [6] อวยพร เรืองตระกูล, “การวิเคราะห์การถดถอยพหุคูณ (multiple regression)” ภาควิชาวิจัยและจิตวิทยาการศึกษา ครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย.
- [7] อุดลย์ ยิ้มงาม, การทำเหมืองข้อมูล (Data Mining) สืบค้นเมื่อ 12 เมษายน 2559, จาก <http://compcenter.bu.ac.th/news-information/data-mining>
- [8] อภิสิทธิ์ บุญญาวารกุล, โรคอ้วน สืบค้นเมื่อ 29 มีนาคม 2559, จาก http://www.healthtoday.net/thailand/disease/disease_134.html

- [9] Chinnapat Kaewchinporn, ต้นไม้ตัดสินใจ (Decision Tree) สืบค้นเมื่อ 30 มีนาคม 2559, จาก http://sasdkmitl09.blogspot.com/2009/07/blog-post_23.html
- [10] Scorpio, *Weka* สำหรับทำ DataMining สืบค้นเมื่อ 11 เมษายน 2559, จาก <http://goo.gl/wsY6Ew>