

การพัฒนาแบบจำลองปัจจัยที่มีผลต่อการเป็นโรคเบาหวานด้วยเทคนิคต้นไม้ตัดสินใจ DEVELOPMENT OF MODEL FOR DIABETES MELLITUS USING DECISION TREE TECHNIQUE.

วนิดา พงษ์สงวน¹, ทิพยา ถินสูงเนิน² และ มาโนช ถินสูงเนิน³

สาขาวิชาวิทยาการคอมพิวเตอร์^{1,2,3} คณะวิทยาศาสตร์และเทคโนโลยี^{1,2,3} มหาวิทยาลัยราชภัฏนครราชสีมา^{1,2,3}

E-mail: nida311662@hotmail.com^{1*}, E-mail: tippayasot@hotmail.com²

บทคัดย่อ

โครงการนี้มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองของปัจจัยที่มีผลต่อการเป็นโรคเบาหวานด้วยเทคนิคเหมืองข้อมูลแบบต้นไม้ตัดสินใจเพื่อช่วยในการวิเคราะห์หาแบบจำลองของปัจจัยที่มีผลต่อการเป็นโรคเบาหวาน ปัจจัยเสี่ยงที่มีผลต่อการเกิดโรคเบาหวานถือเป็นสารสนเทศที่มีความสำคัญที่จะช่วยให้หน่วยงานทางการแพทย์นำไปใช้สำหรับวางแผนกลยุทธ์กลุ่มเป้าหมายในการดำเนินงานวิจัยใช้ข้อมูลผู้เข้ารับบริการที่โรงพยาบาลด้านชุมชน จังหวัดนครราชสีมา ระหว่างปี 2550 – 2555 จำนวนทั้งสิ้น 4,402 ราย แบ่งข้อมูลสำหรับฝึกและทดสอบแบบจำลองออกเป็นร้อยละ 90:10 ตามลำดับ พัฒนาแบบจำลองด้วยอัลกอริทึมเจสึสแปด ซึ่งเป็นเทคนิคแบบต้นไม้ตัดสินใจ ประเมินประสิทธิภาพของแบบจำลองด้วยค่าความแม่นยำ ผลการวิจัยพบว่าแบบจำลองที่พัฒนาให้ประสิทธิภาพที่มีค่าความแม่นยำ 76.14% และสามารถสร้างกฎการจำแนกจากต้นไม้ตัดสินใจทั้งสิ้น 97 กฎ ซึ่งพบว่าปัจจัยเสี่ยงที่อาจก่อให้เกิดโรคเบาหวาน ได้แก่ อายุ เพศ สถานะภาพ ที่อยู่ อาชีพ ประวัติความดันโลหิตเกินมาตรฐาน ประวัติค่าดัชนีมวลกายเกินมาตรฐาน พฤติกรรมการสูบบุหรี่ พฤติกรรมการดื่มสุรา และประวัติครอบครัวเป็นเบาหวาน

คำสำคัญ: โรคเบาหวาน , เหมืองข้อมูล , ต้นไม้ตัดสินใจ , อัลกอริทึมเจสึสแปด

ABSTRACT

The purposes of the research were to develop a diabetes mellitus model with data mining technique using the decision tree for analysis to finding the diabetes risk factors. The risk factors for diabetes are important information that will support medical agencies to strategies planning for disease prevention to direct to group target. On conducting the research, data were collected from patients at Dan Khun Thot Hospital from Nakhon Ratchasima Province during 2007-2012, there were 4,402 cases. The data are separated into 2 sets that training set and testing set were, 90:10 percent, respectively. The modeling using j48 algorithms that a decision tree technique. On evaluating a performance of models, we use an accuracy. The results showed that the performance of model developed has accuracy 76.14% and rules of classification for decision trees have 97 rules. The risk factors that may cause diabetes were age, gender, status, address, occupational, hypertension history, history of body mass index excess, smoking, drinking and family history of diabetes.

Keyword: Diabetes Mellitus, Data Mining, Decision Tree, J48 Algorithm

บทนำ

เบาหวานเป็นโรคเรื้อรัง ที่เป็นปัญหาสำคัญทางด้านสาธารณสุขของโลก รวมทั้งประเทศไทยสถานการณ์เบาหวานปัจจุบัน มีอุบัติการณ์เพิ่มมากขึ้น จากข้อมูลสมาพันธ์เบาหวานนานาชาติ (International diabetes

Federation: IDF) ได้รายงานสถานการณ์ผู้เป็นเบาหวานทั่วโลกแล้ว 285 ล้านคน และได้ประมาณการว่าจะมีจำนวนผู้เป็นเบาหวานทั่วโลกเพิ่มมากกว่า 435 ล้านคน ในปี พ.ศ. 2573 หากไม่มีการดำเนินการในการป้องกันและควบคุมที่มีประสิทธิภาพ สำหรับประเทศไทย ข้อมูลจากสำนักงานนโยบาย

และยุทธศาสตร์ กระทรวงสาธารณสุขพบมีผู้เสียชีวิตจากโรคเบาหวานใน ปี พ.ศ. 2555 ประมาณ 12,100 คน หรือประมาณวันละ 33.2 คน เมื่อพิจารณาระหว่างปี พ.ศ. 2542-2552 พบคนไทยนอนรักษาตัวที่โรงพยาบาลสังกัดกระทรวงสาธารณสุข ด้วยโรคเบาหวาน เพิ่มขึ้น 4.02 เท่า เฉพาะ ปี พ.ศ. 2552 มีผู้ทันนรักษาตัวที่โรงพยาบาลสังกัดกระทรวงสาธารณสุข ด้วยโรคเบาหวาน 558,156 ครั้ง เมื่อแยกการกระจายความชุกออกเป็นรายภาค พบเรียงลำดับจากมากไป น้อย ดังนี้ กรุงเทพฯ , ภาคกลาง , ตะวันออกเฉียงเหนือ , ภาคเหนือ และ ภาคใต้ (ความชุกร้อยละ 9.2 , 7.6 , 7.0 , 5.7 และ 5.0 ตามลำดับ)

จากสถิติข้างต้นนับเป็นปัญหาสำคัญที่ต้องเร่งหาแนวทางแก้ไข ป้องกัน บรรเทา และลดอัตราการเกิดโรคเบาหวาน ซึ่งจะช่วยลดอัตราการตายจากโรคเบาหวานลงได้ เทคโนโลยีด้านเหมืองข้อมูลเป็นตัวเลือกหนึ่งที่เหมาะสมสำหรับช่วยในการวิเคราะห์สร้างแบบจำลองของปัจจัยที่ก่อให้เกิดโรคเบาหวาน สามารถพัฒนาต่อยอดเป็นเครื่องมือสำหรับนำไปสร้างแบบประเมินความเสี่ยงโรคเบาหวานโดยไม่ต้องอาศัยการตรวจเลือด สามารถนำไปใช้ในการสร้างระบบเตือนภัย ทั้งช่วยให้เกิดการปรับเปลี่ยนพฤติกรรมที่เป็นปัจจัยการเกิดโรคเบาหวานได้อย่างตรงเป้าหมาย เนื่องจากเทคนิคการเรียนรู้ของเครื่องจักรเป็นเทคโนโลยีที่สามารถทำงานกับข้อมูลปริมาณมากได้อย่างมีประสิทธิภาพ โดยเทคนิคเหมืองข้อมูลที่มีรูปแบบการแสดงผลที่เข้าใจง่าย การทำงานไม่ซับซ้อน ประมวลผลได้รวดเร็ว และเป็นที่ยอมรับสำหรับงานจำแนกข้อมูล ได้แก่เทคนิคต้นไม้ตัดสินใจ ซึ่งพบจากหลายงานวิจัยที่พยายามศึกษาหาต้นความเสี่ยงต่อโรคเบาหวาน เช่นในงานวิจัยของ Zorman และคณะ (2002) ค้นหาปัจจัยที่เกี่ยวข้องกับการเกิดโรคเบาหวาน โดยอาศัยการผสมผสานระหว่างเทคนิคกฎความสัมพันธ์และเทคนิคต้นไม้ตัดสินใจ ในงานวิจัยของ กิตติพล วิแสง และคณะ (2552) ได้ศึกษาและวิเคราะห์ปัจจัยเสี่ยงของโรคเบาหวานโดยเปรียบเทียบเทคนิค Back propagation Neural Networks, Radial Basis Function Network และตัวแบบ Naïve Bayes เป็นต้น

งานวิจัยนี้จึงได้นำเสนอ การพัฒนาแบบจำลองปัจจัยที่มีผลต่อการเป็นโรคเบาหวานด้วยเทคนิคเหมืองข้อมูลแบบต้นไม้ตัดสินใจ โดยใช้ข้อมูลผู้เข้ารับบริการที่โรงพยาบาลด่านขุนทด จังหวัดนครราชสีมา ระหว่างปี 2550 – 2555 จำนวนทั้งสิ้น 4,402 ราย แบ่งข้อมูลสำหรับฝึกและทดสอบแบบจำลองออกเป็นร้อยละ 90:10 ตามลำดับ พัฒนาแบบจำลองด้วยอัลกอริทึมเจสสิบแปด ซึ่งเป็นเทคนิคแบบต้นไม้ตัดสินใจ ประเมินประสิทธิภาพของแบบจำลอง

ด้วยค่าความแม่นยำ โดยอาศัยโปรแกรมประยุกต์ด้านเหมืองข้อมูล เพื่อค้นหาปัจจัยที่อาจเป็นสาเหตุของการเกิดโรคเบาหวานได้

วัตถุประสงค์การวิจัย

1. เพื่อค้นหาแบบจำลองปัจจัยที่มีผลต่อการเกิดโรคเบาหวาน ของประชาชนในเขตอำเภอด่านขุนทด จังหวัดนครราชสีมา โดยใช้เทคนิคเหมืองข้อมูลแบบต้นไม้ตัดสินใจ
2. เพื่อหาคุณลักษณะที่มีผลต่อการเป็นโรคเบาหวาน ของประชาชนในเขตอำเภอด่านขุนทด จังหวัดนครราชสีมา
3. เพื่อนำเสนอแบบจำลองที่ได้ต่อผู้อำนวยการโรงพยาบาลด่านขุนทด เพื่อใช้เป็นข้อมูลจัดทำกลยุทธ์ในการส่งเสริม กระตุ้นเตือน ให้ความรู้แก่ประชาชน ในเขตอำเภอด่านขุนทด จังหวัดนครราชสีมา

วิธีดำเนินการวิจัย

1. ขั้นตอนการดำเนินการวิจัย

- 1.1 ศึกษาการทำเหมืองข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจ และข้อมูลปัจจัยที่ทำให้เกิดโรคเบาหวาน
- 1.2 การประสานความร่วมมือจากโรงพยาบาลด่านขุนทด ในการขอข้อมูล และเก็บข้อมูลที่จำเป็นต่อการดำเนินงานโครงการ
 - 1) จัดทำหนังสือบันทึกถึงผู้อำนวยการโรงพยาบาลด่านขุนทด เพื่อขอความร่วมมือ
 - 2) ประสานกลุ่มงานประกันสุขภาพ ยุทธศาสตร์และสารสนเทศทางการแพทย์ เพื่อขอความร่วมมือ
 - 3) ประสานศูนย์สารสนเทศเพื่อขอความร่วมมือในการดึงฐานข้อมูล
- 1.3 สืบค้น และเก็บรวบรวมข้อมูล ดังนี้
 - 1) ข้อมูลผู้ป่วยโรคเบาหวานจากเวชระเบียนผู้ป่วยนอก (คลินิกโรคไม่ติดต่อ, คลินิกเบาหวาน)
 - 2) ข้อมูลผู้ป่วยที่ไม่เป็นโรคเบาหวานจากเวชระเบียนผู้ป่วยนอก
- 1.4 ออกแบบฐานข้อมูลเพื่อสำรวจข้อมูลของกลุ่มตัวอย่างในเขตอำเภอด่านขุนทด จังหวัดนครราชสีมา ที่เป็นโรคเบาหวาน และไม่เป็นโรคเบาหวาน
- 1.5 นำเข้าข้อมูล สร้างระบบฐานข้อมูลด้วยระบบจัดการฐานข้อมูลเพื่อจัดเก็บข้อมูลที่ได้จากการสำรวจให้เป็นระเบียบ เป็นมาตรฐานเดียวกัน
- 1.6 จัดทำคลังข้อมูล เพื่อเตรียมสำหรับการทำเหมืองข้อมูล มีขั้นตอนดังนี้

1) รวบรวมข้อมูลจากโรงพยาบาลด้านขุนทด
2) รวมข้อมูลจากเวชระเบียนผู้ป่วยที่ทำ
การบันทึกหลักฐานข้อมูล และข้อมูลที่ตั้งจากฐานข้อมูลจาก
โรงพยาบาลด้านขุนทด ให้เป็นชุดเดียวกัน

3) ทำความสะอาดข้อมูล

4) แปลงข้อมูลให้อยู่ในรูปแบบที่ต้องการ

1.7 พัฒนาแบบจำลองปัจจัยที่มีผลต่อการเป็น
โรคเบาหวาน ด้วยโปรแกรม WEKA 3.7.5

1.8 ทดสอบประสิทธิภาพของแบบจำลอง ทั้ง
วิธีแยกชุดข้อมูล (Split Testing) ชุดฝึก/ชุดทดสอบ และ
เทคนิคแบบไขว้ (k-fold Cross-validation Testing)

1.9 ใช้แบบจำลองที่ได้ ทำนายผลข้อมูลที่สุ่ม
แยกไว้ 400 รายการ

1.10 สรุปผลการดำเนินงานและปัญหาที่พบใน
การดำเนินงานวิจัย

1.11 นำเสนอแบบจำลองที่ได้ ต่อผู้บริหาร
โรงพยาบาลด้านขุนทด จังหวัดนครราชสีมา

2. เครื่องมือการวิจัย

2.1 ระบบบริหารจัดการฐานข้อมูล Microsoft
Access 2007 สำหรับจัดเก็บข้อมูลให้เป็นมาตรฐานเดียวกัน

2.2 โปรแกรมประยุกต์ด้านงานคำนวณ
Microsoft Excel 2007 สำหรับสนับสนุนการเตรียมข้อมูล

2.3 โปรแกรมประยุกต์ด้านงานเหมืองข้อมูล
WEKA 3.7.5 สำหรับวิเคราะห์และพัฒนาแบบจำลอง

2.4 แบบสอบถามสำหรับรวบรวมข้อมูล

3. ประชากรและกลุ่มตัวอย่าง

ข้อมูลที่ใช้สำหรับงานวิจัยนี้มี คือ ข้อมูลกลุ่ม
ตัวอย่างผู้เข้ารับบริการที่โรงพยาบาลด้านขุนทด จังหวัด
นครราชสีมา ระหว่างปี 2550 – 2555 จำนวนทั้งสิ้น 4,402
ราย จำแนกเป็นผู้เข้ารับบริการที่ป่วยเป็นโรคเบาหวาน
จำนวน 1,898 ราย ไม่เป็นโรคเบาหวาน 2,504 ราย

4. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ได้แก่ การรวบรวมข้อมูล เทคนิคการจำแนกแบบ
ต้นไม้ตัดสินใจ อัลกอริทึม j48 การประเมินแบบจำลอง และ
งานวิจัยที่เกี่ยวข้อง ดังนี้

4.1 การรวบรวมข้อมูลสำหรับการวิจัย

ข้อมูลสำหรับงานวิจัยนี้รวบรวมจาก
โรงพยาบาลด้านขุนทด จังหวัดนครราชสีมา แสดงตัวอย่างได้
ดังตารางที่ 1 โดยรวบรวมข้อมูลจาก 2 แหล่งได้แก่

1) จากแฟ้มประวัติผู้ป่วย (OPD Cards)
เป็นข้อมูลที่ยังไม่มีการจัดเก็บลงสื่อดิจิทัล

2) จากฐานข้อมูลโรงพยาบาลด้านขุนทด
เป็นข้อมูลที่ถูกจัดเก็บในระบบบริหารจัดการฐานข้อมูลแล้ว
ซึ่งรวบรวมโดยขอความอนุเคราะห์จากศูนย์สารสนเทศ กลุ่ม
งานประกันสุขภาพ ยุทธศาสตร์และสารสนเทศทางการแพทย์
โรงพยาบาลด้านขุนทด ซึ่งข้อมูลที่ถูกจัดเก็บจำแนกเป็น
กลุ่มได้ 3 กลุ่มหลัก ดังต่อไปนี้

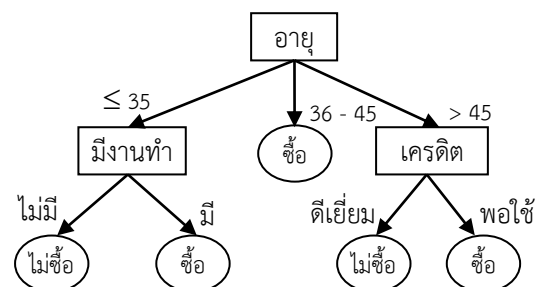
1) ข้อมูลส่วนบุคคล ได้แก่ เพศ อายุ
ภูมิถิ่นอาศัย อาชีพ สถานะภาพ ประวัติครอบครัว และผลการ
วินิจฉัยทางการแพทย์

2) ข้อมูลพฤติกรรมกรรมการดำเนินชีวิต ได้แก่
ข้อมูลพฤติกรรมการสูบบุหรี่ และพฤติกรรมการดื่มสุรา

3) ข้อมูลทางกายภาพ ได้แก่ ส่วนสูง
น้ำหนัก และความดันโลหิต ระหว่างปี 2550–2555

4.2 เทคนิคต้นไม้ตัดสินใจ

เทคนิคต้นไม้ตัดสินใจ เป็นเทคนิคที่ใช้ในงาน
ด้านเหมืองข้อมูลประเภทการจำแนกข้อมูลที่มีจุดเด่นคือ
แสดงผลลัพธ์เป็นผังต้นไม้ ซึ่งเป็นวิธีที่ได้รับความนิยม
เนื่องจากอัลกอริทึมไม่ซับซ้อน ตีความและเข้าใจลักษณะของ
รูปแบบข้อมูลได้ง่าย (ชนวิวัฒน์ ศรีสุอาน, 2550) การ
พยากรณ์ในรูปของต้นไม้ตัดสินใจเป็นการเรียนรู้แบบมีผู้สอน
(Supervised Learning) หมายถึงการสร้างแบบจำลองจาก
ข้อมูลที่รู้กลุ่มที่แท้จริงก่อนล่วงหน้า รูปแบบของต้นไม้
ประกอบด้วยโหนด โดยโหนดแรกเรียกว่ารากโหนด จากราก
โหนดจะแยกออกเป็นโหนดลูก ที่ระดับสุดท้ายจะเรียกว่า
โหนดใบ ซึ่งจะแสดงคลาสที่เป็นผลลัพธ์ที่สามารถแยกแยะได้
แสดงได้ดังภาพที่ 1



ภาพที่ 1 ตัวอย่างต้นไม้ตัดสินใจซื้อโทรศัพท์มือถือ

จากภาพที่ 1 ตัวอย่างต้นไม้ตัดสินใจซื้อ
โทรศัพท์มือถือ ซึ่งแสดงแบบจำลองการตัดสินใจ ซื้อ/ไม่ซื้อ
โทรศัพท์มือถือ สรุปได้ว่าคนที่ตัดสินใจซื้อโทรศัพท์มือถือ คือ
“ผู้มีอายุ ระหว่าง 36-45 ปี” หรือ “ผู้มีอายุต่ำกว่า 36 ปี
และมีงานทำ” หรือ “ผู้มีอายุมากกว่า 45 ปี และมีเครียด
พอใช้”

4.3 อัลกอริทึม J48

อัลกอริทึม j48 (ในโปรแกรม Weka) หรือ อัลกอริทึม C4.5 เป็นอัลกอริทึมอีกหนึ่งที่ใช้สร้างต้นไม้ตัดสินใจที่พัฒนาโดย Ross Quinlan ซึ่งปรับปรุงจาก อัลกอริทึม ID3 (Dunham, 2006; Sharma and Sahni, 2011; Bhargava et al., 2013) โดย C4.5 จะสร้างผังต้นไม้ตัดสินใจมาจากกลุ่มของข้อมูลชุดฝึกที่ใช้หลักการของ Information Entropy (Quinlan, 1996) ในการเลือกโหนดโดยใช้ค่าความถูกต้องของคุณลักษณะแต่ละตัวเพื่อใช้แยกเส้นทางตัดสินใจจะพิจารณาจากค่า Normalized Information Gain

ในการใช้งานเทคนิคการจำแนกแต่เทคนิคเราไม่สามารถบอกได้ว่าเทคนิคใดดีที่สุด ดังนั้นจึงควรทดลองสร้างแบบจำลองหลายครั้งแล้วนำมาเปรียบเทียบกันว่าแบบจำลองใดให้ประสิทธิภาพได้ดีที่สุด (พวงทิพย์ แท่นแสง, 2550) สำหรับอัลกอริทึม J48 มีการทำงาน (Patil and Sherekar, 2013; Dunham, 2006) ดังภาพที่ 2

```

INPUT:
    D    //Training Data
OUTPUT
    T    //Decision Tree
DTBUILD (*D) {
    T=φ;
    T=Create root node and label with splitting attributes;
    T=Add arc to root node for each split predicate and label;
    For each arc do
        D=Database created by applying splitting predicate to D;
        If stopping point reached for this path, then
            T'=Create leaf node and label with appropriate class;
        Else
            T'=DTBUILD(D);
        T=add T' to arc;
    }
  
```

ภาพที่ 2 อัลกอริทึม J48

จากภาพที่ 2 การทำงานของอัลกอริทึม J48 จะทำงานวนซ้ำในการเลือก Attributes ที่เหมาะสมให้เป็นโหนดตัดสินใจไปเรื่อย ๆ จนกระทั่งทุกโหนดสุดท้ายของแต่ละเส้นทางจะเป็นโหนดใบ ซึ่งหมายถึงคลาสที่เหมาะสมสำหรับเส้นทางการตัดสินใจแต่ละเส้นทาง

4.4 วิธีการทดสอบแบบจำลอง งานวิจัยนี้ใช้

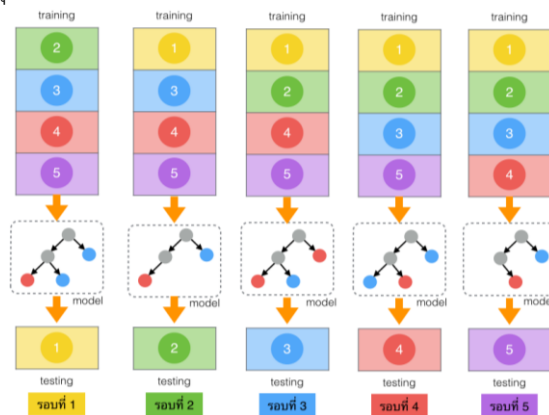
วิธีการทดสอบแบบจำลอง 2 วิธีดังนี้

1) การทดสอบด้วยการแยกชุด

วิธีทดสอบแบบแยกชุด เป็นการแบ่งข้อมูลด้วยการสุ่มออกเป็น 2 ส่วน ส่วนแรกคือชุดสำหรับเรียนรู้เพื่อสร้างแบบจำลอง เรียกว่า ชุดฝึก (Training Set) อีกส่วนหนึ่งไว้สำหรับทดสอบแบบจำลอง เรียกว่า ชุดทดสอบ (Testing Set) เช่น 70:30 หมายถึงชุดฝึก 70% ชุดทดสอบ 30% โดยประสิทธิภาพของการสร้างแบบจำลองด้วยการทดสอบแบบวิธีนี้มักขึ้นกับความสามารถในการสุ่มแยกข้อมูล ถ้าการสุ่มข้อมูลชุดทดสอบได้คล้ายกับข้อมูลชุดฝึกผลประเมินประสิทธิภาพที่ได้จะดี แต่ข้อดีของวิธีการนี้คือใช้เวลาอันน้อยซึ่งเหมาะกับชุดข้อมูลที่มีขนาดใหญ่

2) การทดสอบแบบไขว้

วิธีทดสอบแบบไขว้เป็นวิธีการวัดประสิทธิภาพที่การแบ่งข้อมูลออกเป็นหลายส่วน (แทนด้วย k-fold) เช่น 10-fold cross-validation คือ การแบ่งข้อมูลออกเป็น 10 ส่วน แต่ละส่วนมีจำนวนข้อมูลเท่ากัน สร้างแบบจำลองจากข้อมูล 9 ชุด แล้วใช้ 1 ชุดที่เหลือสำหรับทดสอบ วนซ้ำการทำงานนี้จนกระทั่งข้อมูลทุกชุดถูกใช้เป็นชุดทดสอบ ตัวอย่างดังภาพที่ 3



ภาพที่ 3 การทดสอบแบบจำลองด้วยวิธีไขว้ 5-fold

(ที่มา: เอกสิทธิ์ พัทธวงศ์ศักดิ์, 2557)

4.5 การประเมินประสิทธิภาพ

การประเมินประสิทธิภาพของแบบจำลอง มีหลายมาตรวัดให้เลือกใช้งาน สำหรับงานวิจัยนี้ค่าความแม่นยำ (Accuracy) ซึ่งหาได้จากการพิจารณาตาราง Confusion Matrix ดังแสดงในตารางที่ 1

ตารางที่ 1 Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

จากตารางที่ 1 Confusion Matrix คือ การประเมินผลลัพธ์การทำนายเปรียบเทียบกับผลลัพธ์จริง โดยพิจารณาค่า ดังนี้

True Positive (TP) คือ จำนวนที่ทายคลาส Positive ถูก
True Negative (TN) คือ จำนวนที่ทายคลาส Negative ถูก
False Positive (FP) คือ จำนวนที่ทายคลาส Positive ผิด
False Negative (FN) คือ จำนวนที่ทายคลาส Negative ผิด

สำหรับมาตรวัดค่าความแม่นยำ (Accuracy) คือ ค่าที่บอกว่าการทำนาย ทำนายได้แม่นยำแค่ไหน สามารถคำนวณได้ดังนี้

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

4.6 งานวิจัยที่เกี่ยวข้อง

Zorman และคณะ (2002) ได้ศึกษาวิจัยเรื่อง Mining Diabetes Database with Decision Trees and Association Rules เป็นการศึกษาเพื่อหาวิธีการที่จะผลิตกฎสำหรับช่วยวินิจฉัยโรคเบาหวานที่มีความกระชับมากกว่าระหว่างเทคนิคต้นไม้ตัดสินใจกับเทคนิคกฎความสัมพันธ์ โดยข้อมูลในการวิจัยคือฐานข้อมูลโรคเบาหวานจาก Osaka Medical College Hospital ผลการวิจัยพบว่ากฎที่ได้จากต้นไม้ตัดสินใจจะมีจำนวนน้อยกว่ากระชับกว่ากฎที่ได้จากกฎความสัมพันธ์

กิตติพล วิแสง และคณะ (2552) ได้ศึกษาและวิเคราะห์ปัจจัยเสี่ยงของโรคเบาหวาน โดยเปรียบเทียบการใช้เทคนิค Back propagation Neural Networks (BPNN), Radial Basis Function Network (RBFN) และ Naïve Bayes โดยอ้างอิงจากผลการวินิจฉัยทางการแพทย์เพื่อวัดความถูกต้องของแบบจำลองที่ได้จาก 3 เทคนิค ซึ่งผลการวิจัยพบว่า แบบจำลองที่ได้จากเทคนิค BPNN ให้ผลลัพธ์ที่มีความถูกต้องมากที่สุด โดยให้ค่า MAE เท่ากับ 0.0045 ค่า RMSE เท่ากับ 0.0126 และประสิทธิภาพการจำแนกกลุ่มสมาชิกข้อมูลเปรียบเทียบผลวินิจฉัยทางการแพทย์เท่ากับ 100 % (เมื่อทดสอบกับข้อมูลปัจจัยเสี่ยงชุดที่ 2) และปัจจัยเสี่ยงในการเกิดโรคเบาหวาน คือ BMI, FDM, AGE, HT, OR และ SEX ตามลำดับ

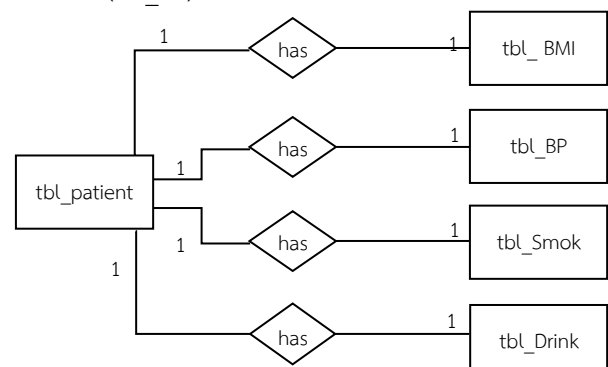
กิตติศักดิ์ สุมาลย์ (2555) นำเสนอวิธีการวิเคราะห์ข้อมูลประชาชนในเขตเทศบาลเมืองจังหวัดบุรีรัมย์ ปี 2555 จำนวน 1,071 ราย ด้วยเทคนิคกฎความสัมพันธ์ สำหรับการศึกษาความสัมพันธ์ของโรคและใช้เทคนิคต้นไม้ตัดสินใจสำหรับการศึกษาคัดกรองสุขภาพเบื้องต้นของประชาชน โดยแบ่งเป็น กลุ่มปกติ กลุ่มเสี่ยง และกลุ่มป่วย ผลการวิจัยพบว่าการกฎความสัมพันธ์ที่ได้จากอัลกอริทึม FP

Growth ให้คำตอบที่เหมาะสมกว่า อัลกอริทึม Apriori และสำหรับการคัดกรองด้วยเทคนิคต้นไม้ตัดสินใจพบว่าแบบจำลองที่ได้ให้ค่าความถูกต้อง (Precision) เท่ากับ 88.60% ค่าระลึก (Recall) เท่ากับ 89.20% ค่าความเที่ยงหรือประสิทธิภาพโดยรวม (F-Measure) เท่ากับ 88.80% และได้จำนวนกฎที่ผ่านการตรวจสอบจากผู้เชี่ยวชาญพบว่าเหลือกฎความสัมพันธ์ 42 กฎ

ผลการวิจัย

1. ผลการเตรียมข้อมูลก่อนการประมวลผล

1.1 ผลการออกแบบฐานข้อมูล ผู้วิจัยได้ดำเนินการวิเคราะห์และออกแบบฐานข้อมูลโดยมีจำนวนตารางข้อมูลทั้งหมด 5 ตาราง ได้แก่ ตารางรายละเอียดข้อมูลทั่วไป (tbl_patient) ตารางรายละเอียดข้อมูลพฤติกรรมการดื่มสุรา (tbl_drink) ตารางรายละเอียดข้อมูลพฤติกรรมการสูบบุหรี่ (tbl_smok) ตารางรายละเอียดข้อมูลทางกายภาพ (tbl_BMI) และ ตารางรายละเอียดข้อมูลทางกายภาพ (tbl_BP) โดยมีความสัมพันธ์แสดงดังภาพที่ 4



ภาพที่ 4 แผนภาพความสัมพันธ์ของข้อมูล

จากภาพที่ 4 แผนภาพแสดงความสัมพันธ์ของข้อมูลประกอบด้วย 5 ตาราง โดยตาราง tbl_patient มีความสัมพันธ์แบบ 1:1 กับตาราง tbl_BML, tbl_BP, tbl_Smok และ tbl_Drink

1.2 การทำความสะอาดข้อมูล สำหรับข้อมูลบางรายการหรือบางคุณลักษณะอาจจัดเก็บค่าที่ไม่คาดคิด (Noisy Data) หรือบางค่าขาดหายไป (Missing Values) ซึ่งปัญหานี้อาจส่งผลกระทบต่อประสิทธิภาพของแบบจำลอง จำเป็นต้องมีการทำความสะอาดข้อมูล ดังนั้นงานวิจัยนี้จัดการกับข้อมูลที่ไม่คาดคิดด้วยการตัดรายการนั้นทิ้งสำหรับค่าที่หายไปแก้ปัญหาโดยการแทนข้อมูลด้วย “?” ซึ่งเป็นตัวเลือกหนึ่งในโปรแกรม Weka

1.3 การแปลงข้อมูล สำหรับข้อมูลที่รวบรวมได้มีบางคุณลักษณะที่ยังไม่เหมาะสมสำหรับการประมวลผล ดังนั้นงานวิจัยนี้จึงได้แปลงข้อมูลให้เหมาะสมดังตารางที่ 2

ตารางที่ 2 การแปลงข้อมูล

คุณลักษณะ	ค่าที่รวบรวมได้	แทนด้วย
อายุ	<=43, >43	ตัวเลขอายุ
ที่อยู่	ในเขตอำเภอตำบลขุนทด นอกเขตอำเภอตำบลขุนทด	1 0
BMI (ดัชนีมวลกาย)	ดัชนีมวลกายเกินมาตรฐาน น้อยกว่า 2 ครั้ง มากกว่าหรือเท่ากับ 2 ครั้ง	No Yes
BP	เกินมาตรฐาน น้อยกว่า 5 ครั้ง เกินมาตรฐาน มากกว่าหรือเท่ากับ 5 ครั้ง	No Yes

จากตารางที่ 2 การแปลงข้อมูล มีบางคุณลักษณะที่จำเป็นต้องแปลงข้อมูลให้เหมาะสมต่อการประมวลผลคือ อายุ ที่อยู่ ดัชนีมวลกาย (BMI) และความดันเลือดเกินมาตรฐาน (BP)

1.4 การเตรียมชุดข้อมูล งานวิจัยนี้นำข้อมูลจากตารางในฐานข้อมูลมาหลอมรวมกันไว้เป็นคลังข้อมูล โดยเตรียมชุดข้อมูลสำหรับทดสอบจำนวน 11 ชุด แบ่งเป็นข้อมูลพื้นฐาน ข้อมูลพื้นฐานและความดันโลหิต ข้อมูลพื้นฐานและดัชนีมวลกาย ข้อมูลพื้นฐานและพฤติกรรม

การสูบบุหรี่ ข้อมูลพื้นฐานและพฤติกรรมการดื่มสุรา ข้อมูลทางกายภาพและพฤติกรรมการดื่ม-การสูบบุหรี่ ข้อมูลทางกายภาพ-พฤติกรรมการดื่ม-การสูบบุหรี่และประวัติครอบครัว ข้อมูลทั้งหมด (เว้น สถานภาพและที่อยู่) ข้อมูลทั้งหมด โดยข้อมูลทั้ง 11 ชุดถูกจัดเก็บเป็นชนิด csv

ดังตารางที่ 3

2. ผลการออกแบบการทดลอง

ในการพัฒนาแบบจำลองประกอบด้วย 2 กิจกรรม ได้แก่ การสร้าง และทดสอบแบบจำลอง ซึ่งงานวิจัยนี้ออกแบบการทดลองแบ่งเป็น 2 แนวทางดังนี้

2.1 การทดสอบด้วยวิธีแบบแยกชุด งานวิจัยนี้แยกชุดข้อมูลไว้ 3 แบบ คือ 70:30 80:20 และ 90:10 ตามลำดับ

2.2 การทดสอบด้วยวิธีแบบไขว้ งานวิจัยนี้กำหนดการทดสอบไขว้ไว้ 2 แบบ ได้แก่ 5-fold และ 10-fold Cross-validation

ตารางที่ 3 ตัวอย่างชุดข้อมูลเตรียมนำเข้าสู่การประมวลผล ชุดที่ 11 ข้อมูลทั้งหมด

Gender	Age	Addr	Occu	Status	BP	BMI	Smoking	Drinking	F_DM	DM
male	34	1	government official	single	No	No	no smoking	no drinking	Yes	No
male	38	1	government official	single	No	Yes	no smoking	no drinking	Yes	No
male	46	1	individual	single	Yes	Yes	smoking	drinking	Yes	Yes
female	39	1	agriculturist	marry	Yes	No	no smoking	no drinking	No	Yes
female	50	1	agriculturist	marry	No	No	ever	drinking	No	Yes
female	79	1	agriculturist	marry	Yes	Yes	no smoking	ever	Yes	Yes

จากตารางที่ 3 ตัวอย่างชุดข้อมูลเตรียมนำเข้าสู่การประมวลผล ชุดที่ 11 ซึ่งประกอบด้วยคุณลักษณะ เพศ (Gender) อายุ (Age) ที่อยู่ (Addr) อาชีพ (Occu) สถานะภาพสมรส (Status) ความดันเลือดเกินมาตรฐาน (BP) ดัชนีมวลกาย (BMI) พฤติกรรมการสูบบุหรี่ (Smoking) พฤติกรรมการดื่มสุรา (Drinking) ประวัติครอบครัวเป็นเบาหวาน (F_DM) และผลการวินิจฉัยโรคเบาหวาน (DM) ตามลำดับ

3. ผลการพัฒนาแบบจำลอง

ผู้วิจัยได้ดำเนินการพัฒนาแบบจำลองของปัจจัยที่มีผลต่อการเป็นโรคเบาหวานด้วยเทคนิคเหมืองข้อมูลแบบต้นไม้ตัดสินใจ ตามขั้นตอนการวิจัยได้ผลการดำเนินงานดังตารางที่ 4

ตารางที่ 4 ผลการพัฒนาแบบจำลอง

ชุดที่	Accuracy				
	70:30	80:20	90:10	5-fold	10-fold
1	63.44	64.77	63.86	64.49	65.11
2	65.18	64.89	63.86	66.72	67.15
3	65.41	66.93	65.00	66.29	66.24
4	65.18	67.95	66.36	66.31	67.33
5	64.57	67.84	65.68	65.97	66.31
6	70.70	71.02	72.73	70.92	72.99
7	63.29	63.30	63.18	65.61	66.63
8	67.22	67.16	64.55	65.38	65.88
9	67.30	67.27	65.91	68.33	69.13
10	70.25	73.41	75.68	71.42	73.42
11	71.23	73.52	76.14	71.29	73.35

จากตารางที่ 4 ผลการพัฒนาแบบจำลอง ซึ่งแสดงผลเปรียบเทียบค่าความแม่นยำที่ได้จากวิธีการที่ออกแบบไว้ 5 วิธีได้แก่ วิธีทดสอบแบบแยกชุด 70:30, 80:20 และ 90:10 วิธีทดสอบแบบไขว้ 5-fold และ 10-fold Cross-validation ซึ่งผลการพัฒนาพบว่าวิธีแยกชุด 70:30, 80:20 และ 90:10 ให้ค่าความแม่นยำสูงสุดเท่ากับ 71.23%, 73.52% และ 76.14% ตามลำดับ ที่ข้อมูลชุดที่ 11 เหมือนกัน สำหรับวิธีแบบไขว้ 5-fold และ 10-fold Cross-validation ให้ค่าสูงสุดเท่ากับ 71.42% และ 73.42% ตามลำดับ ที่ข้อมูลชุดที่ 10 เหมือนกัน

4. ผลการใช้แบบจำลองทำนายผล

จากผลการพัฒนาแบบจำลอง เมื่อเปรียบเทียบวิธีการทดสอบทั้ง 5 วิธีพบว่าวิธีที่ให้ค่าความแม่นยำสูงสุดคือวิธีทดสอบแบบแยกชุด 90:10 ที่ให้ค่าเท่ากับ 76.14% ดังนั้นเมื่อนำแบบจำลองที่ให้ค่าความแม่นยำสูงสุด ใช้ทำนายข้อมูลชุดใหม่ที่เตรียมไว้ 400 รายการ ผลการทำนายพบว่าเป็นแบบจำลองที่ได้ ทำนายถูกต้อง 315 รายการ คิดเป็น 78.75% และทายผิด 85 รายการ คิดเป็น 21.25%

สรุปและอภิปรายผลการวิจัย

จากผลการพัฒนาแบบจำลองของปัจจัยที่มีผลต่อการเป็นโรคเบาหวานด้วยเทคนิคเหมืองข้อมูลแบบต้นไม้ตัดสินใจ อัลกอริทึม J48 เพื่อช่วยในการวิเคราะห์หาแบบจำลองของปัจจัยที่มีผลต่อการเป็นโรคเบาหวาน โดยใช้ข้อมูลจำนวน 4,402 รายการ รวบรวมจากโรงพยาบาลด้านขุนทด จังหวัดนครราชสีมา สามารถสรุปผลการวิจัยได้ดังนี้

1. การเตรียมข้อมูลสำหรับพัฒนาแบบจำลองงานวิจัยนี้ออกแบบข้อมูล 11 ชุด แบ่งเป็น ข้อมูลพื้นฐาน ข้อมูลพื้นฐานและความดันโลหิต ข้อมูลพื้นฐานและดัชนีมวลกาย ข้อมูลพื้นฐานและพฤติกรรมกรรมการสูบบุหรี่ ข้อมูลพื้นฐานและพฤติกรรมกรรมการดื่มสุรา ข้อมูลทางกายภาพและพฤติกรรมกรรมการดื่ม-การสูบบุหรี่ ข้อมูลทางกายภาพ-พฤติกรรมกรรมการดื่ม-การสูบบุหรี่และประวัติครอบครัว ข้อมูลทั้งหมด (เว้น สถานภาพและที่อยู่) และข้อมูลทั้งหมด

2. จากผลการเปรียบเทียบวิธีการทดสอบทั้ง 5 วิธีสรุปได้ว่าวิธีที่ให้ค่าความแม่นยำสูงสุดคือ วิธีทดสอบแบบแยกชุด 90:10 คือแบ่งชุดข้อมูลฝึก 90% และชุดข้อมูลสำหรับทดสอบ 10% ซึ่งให้ค่าความแม่นยำสูงสุดเท่ากับ 76.14%

3. จากผลการวิจัยข้อมูลที่ให้ ประสิทธิภาพแบบจำลองสูงสุดคือข้อมูลชุดที่ 11 สรุปกฎการจำแนกที่ได้จากต้นไม้การตัดสินใจทั้งหมด 97 กฎ และสามารถสรุปปัจจัยเสี่ยงที่ก่อให้เกิดโรคเบาหวาน ได้แก่ อายุ เพศ สถานภาพ และอาชีพ ประวัติความดันโลหิตเกินมาตรฐาน ประวัติค่าดัชนีมวลกายเกินมาตรฐาน พฤติกรรมการสูบบุหรี่ พฤติกรรมการดื่มสุรา

และประวัติครอบครัวเป็นเบาหวาน ซึ่งมีปัจจัยบางตัวสอดคล้องกับทฤษฎี การวิเคราะห์ปัจจัยเสี่ยงของโรคเบาหวาน โดย กิตติพล วิแสง สิริภัทร เชี่ยวชาญวัฒนา และคำณ สุนด์ิ มหาวิทยาลัยขอนแก่น ปี 2552 ที่ได้วิเคราะห์ปัจจัยเสี่ยงในการเกิดโรคเบาหวาน คือ ดัชนีมวลกาย ประวัติครอบครัวเป็นเบาหวาน อายุ ความดันโลหิต ความยาวรอบเอว และ เพศ

4. จากผลการนำแบบจำลองที่ให้ค่าความแม่นยำสูงสุดใช้ทำนายข้อมูลชุดใหม่ที่เตรียมไว้ 400 รายการ สรุปได้ว่าแบบจำลองมีความแม่นยำ 78.75%

ข้อเสนอแนะ

จากผลการวิจัยจะได้ปัจจัยที่ส่งผลต่อการเกิดโรคเบาหวาน ที่มีความแม่นยำ เพื่อให้เกิดประโยชน์สูงสุดในการนำผลไปใช้งาน หรือการทำวิจัยครั้งต่อไป ควรมีการ

1. นำแบบจำลองที่ได้ไปพัฒนาเป็นระบบสนับสนุนการตัดสินใจการประเมินความเสี่ยงการเกิดโรคเบาหวานสำหรับประชาชนเพื่อให้ง่ายต่อการตรวจสอบความเสี่ยงการเกิดโรคเบาหวาน

2. ควรมีการทดสอบในประชากรที่มากขึ้นและควรศึกษาติดตามประชากรกลุ่มเสี่ยงในระยะยาวต่อไป

3. เลือกใช้ข้อมูลหรือปัจจัยอื่นเพิ่มขึ้นเพื่อค้นหาความสัมพันธ์ที่อาจซ่อนอยู่ เพื่อเพิ่มประสิทธิภาพของแบบจำลองให้เกิดความแม่นยำมากขึ้น

4. เลือกเทคนิคการทำเหมืองข้อมูลแบบใหม่เพื่อปรับปรุงประสิทธิภาพของการทำนายผลที่แม่นยำมากขึ้น

5. เลือกอัลกอริทึมที่มีประสิทธิภาพในการปรับปรุงข้อมูล เพื่อทดลองหาวิธีการปรับปรุงข้อมูลที่มีประสิทธิภาพมากยิ่งขึ้น

กิตติกรรมประกาศ

ขอขอบพระคุณ คณาจารย์ และนักศึกษาหลักสูตรวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครราชสีมา ที่ให้การสนับสนุนการทำวิจัย โรงพยาบาลด้านขุนทด จังหวัดนครราชสีมา ที่ให้ความอนุเคราะห์ข้อมูลสำหรับดำเนินงานวิจัย และร่วมติดตามประเมินผลการวิจัยในครั้งนี้

เอกสารอ้างอิง

กิตติพล วิแสง, สิริภัทร เชี่ยวชาญวัฒนา และคำณ สุนด์ิ. (2552). การวิเคราะห์ปัจจัยเสี่ยงต่อโรคเบาหวาน. วิทยานิพนธ์ปริญญาโท. เทคโนโลยีสารสนเทศ. มหาวิทยาลัยขอนแก่น

- กิตติศักดิ์ สุมาลย์. การคัดกรองสุขภาพเบื้องต้นโดยใช้เทคนิคเหมืองข้อมูล [สารนิพนธ์วิทยาศาสตร์มหาบัณฑิต]. กรุงเทพฯ: มหาวิทยาลัยธุรกิจบัณฑิต; 2555. หน้า 1
- ชนวัฒน์ ศรีสอาน. (2550). **ฐานข้อมูล คลังข้อมูลและเหมืองข้อมูล**. กรุงเทพมหานคร : บริษัท เอเชียดิจิตอล จำกัด. ประเทศไทย.
- พวงทิพย์ แท่นแสง. (2550). **การทดสอบประสิทธิภาพการทำงานของอัลกอริทึมไมนิงกฎสำหรับจำแนก.วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต บัณฑิตวิทยาลัย สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ**.
- เอกสิทธิ์ พชรวงศ์ศักดิ์. 2557. **การแบ่งข้อมูลเพื่อทดสอบประสิทธิภาพของโมเดล**. สืบค้นจาก: <http://dataminingtrend.com/2014/data-mining-techniques/cross-validation/>.
- Bhargava, N., Sharma, G., Bhargava, R., and Mathuria, M. (2013). **Decision tree analysis on j48 algorithm for data mining**. Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, 3(6).
- Dunham, M. H. (2006). **Data mining: Introductory and advanced topics**. Pearson Education India.
- Patil, T. R. and Sherekar, S. S. (2013). **Performance analysis of Naive Bayes and J48 classification algorithm for data classification**. International Journal of Computer Science and Applications, 6(2), 256-261.
- Quinlan, J. R. (1993). **C4.5: Programs for Machine Learning**. Morgan Kaufmann Publishers.
- Sharma, A. K. and Sahni, S. (2011). **A comparative study of classification algorithms for spam email data analysis**. International Journal on Computer Science and Engineering, 3(5), 1890-1895.
- Zorman, M., Masuda, G., Kokol, P., Yamamoto and R., Stiglic, B. (2002). **Mining Diabetes Database with Decision Trees and Association Rules**. In Computer-Based Medical Systems, 2002. (CBMS 2002). Proceedings of the 15th IEEE Symposium, 134-139.