รายงานวิจัยฉบับสมบูรณ์

การประยุกต์ใช้โมเดลเครือข่ายแบบเบย์ด้วยวิธีการเรียนรู้ของเครื่อง เพื่อวิเคราะห์ความเสี่ยงการเป็นโรคไม่ติดต่อจากข้อมูลการสำรวจสุขภาพประชาชนไทยโดยการตรวจร่างกาย

Applying Bayesian Network Model for Chronic Disease Risk Analysis by Machine Learning Method from National Health Examination Survey in Thailand

จัดทำโดย

ผศ.ดร.กนกกรรณ์ สิ้โรจนาประภา

ภาควิชาสถิติ คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

งานวิจัยได้รับการสนับสนุนจาก ทุนอุดหนุนการวิจัย

ประเภท เงินอุดหนุนทั่วไป (งบประมาณเงินรายได้) ประจำปี 2559

คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

รายงาน

โครงการวิจัยเรื่อง

การประยุกต์ใช้โมเดลเครือข่ายแบบเบย์ด้วยวิธีการเรียนรู้ของเครื่อง เพื่อวิเคราะห์ความเสี่ยงการเป็นโรคไม่ติดต่อจากข้อมูลการสำรวจสุขภาพประชาชนไทยโดยการตรวจร่างกาย

Applying Bayesian Network Model for Chronic Disease Risk Analysis by Machine Learning Method from National Health Examination Survey in Thailand

จัดทำโดย

ผศ.ดร.กนกกรรณ์ ลี้โรจนาประภา

ภาควิชาสถิติ คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

งานวิจัยได้รับการสนับสนุนจาก ทุนอุดหนุนการวิจัย

ประเภท เงินอุดหนุนทั่วไป (งบประมาณเงินรายได้)

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ประจำปี 2559

สารบัญ

สารบัญตาราง

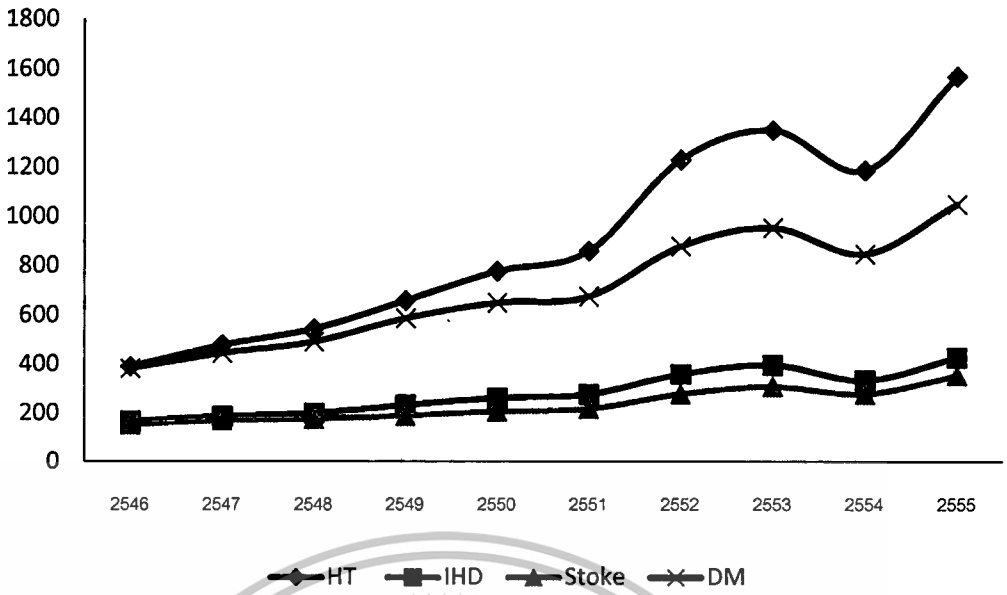
สารบัญรูป

บทที่ 1 บทนำ

งานวิจัยนี้ได้นำความรู้ในการทำโมเดลเพื่อนำมาใช้ในการอธิบายและทำความเข้าใจเกี่ยวกับภาวการณ์เกิดโรคไม่ติดต่อที่สำคัญ โดยมีแนวคิด เป้าหมาย และขอบเขตการศึกษาดังที่จะได้กล่าวไว้ในบทนี้

1.1 ความสำคัญ และที่มาของปัญหา

ประเทศไทยกำลังก้าวสู่ยุคสังคมผู้สูงอายุ ในอีกไม่เกิน 10 ปีจะมีสัดส่วนประชากรอายุ 65 ปีขึ้นไปเพิ่มสูงถึงร้อยละ 14 และต่อจากนั้นอีกไม่เกิน 10 ปี ไทยจะเป็น "สังคมสูงวัยระดับสุดยอด" ในปี พ.ศ. 2575 เมื่อประชากรอายุ 65 ปีขึ้นไปเพิ่มขึ้นถึงร้อยละ 20 (Manager online, 2556) สิ่งที่ตามมาก็คือทั้งภาครัฐบาลและประชาชนผู้ที่กำลังก้าวสู่วัยกลางคนต้องเตรียมรับกับการรักษาจากโรคภัยไข้เจ็บที่มักจะเกิดกับผู้สูงอายุโดยเฉพาะอย่างยิ่งโรคไม่ติดต่อเรื้อรัง โรคไม่ติดต่อเรื้อรัง (Chronic disease) ที่สำคัญประกอบด้วย โรคหัวใจขาดเลือด (IHD) โรคความดันโลหิตสูง (HT) โรคเบาหวาน (DM) โรคหลอดเลือดสมองใหญ่ หรือ อัมพฤกษ์ อัมพาต (Stoke) มีแนวโน้มสูงขึ้นและนำไปสู่อัตราการเสียชีวิตของผู้ป่วยในกลุ่มนี้ที่เพิ่มสูงขึ้น (อมรา ทองหงษ์, กมลชนก เทพสิทธา, & ภาคภูมิ จงพิริยะอนันต์, 2556) ถึงแม้ปัจจุบันยังไม่ได้ก้าวสู่ยุกต์สังคมผู้สูงอายุอย่างเต็มตัวแต่จากรายงานประจำปี 2556 ของสำนักงานโรคไม่ติดต่อ (2556) แสดงสถิติอัตราผู้ป่วยในด้วยโรคไม่ติดต่อที่สำคัญทั้ง 4 โรคดังกล่าว (ทั้งประเทศ) ปี พ.ศ. 2546–2555 มีแนวโน้มสูงขึ้น นอกจากนี้ยังพบว่าสถานการณ์การป่วยในปี 2555 ต่อประชากร 100,000 คน ป่วยด้วยโรคความดันโลหิตสูงสูงสุด (HT) มากที่สุด คิดเป็น 1,570.6 ต่อประชากร 100,000 คน รองลงมาคือเบาหวาน (DM) คิดเป็น 1,050.1 ต่อประชากร 100,000 คน โรคหัวใจขาดเลือด (IHD) คิดเป็น 427.5 ต่อประชากร 100,000 คนและโรคหลอดเลือดสมองใหญ่ หรือ อัมพฤกษ์ อัมพาต (Stoke) คิดเป็น 354.5 ต่อประชากร 100,000 คน (ดังรูป 1-1)



ที่มา: รายงานประจำปี พ.ศ. 2556 (สำนักโรคไม่ติดต่อ กรมควบคุมโรค, 2556)

รูป 1-1 อัตราผู้ป่วยในต่อประชากรแสนคนจำแนกตามโรคไม่ติดต่อเรื้อรังที่สำคัญ ปี พ.ศ. 2546-2555

เพื่อแก้ไขปัญหาด้านการแพทย์และสาธารณสุขอย่างยั่งยืนกระทรวงสาธารณสุขจึงมุ่งเน้นนโยบายเพื่อลดอัตราการเกิดโรคดังกล่าวอย่างเป็นระบบ เป็นที่ทราบกันดีว่า 'การป้องกันย่อมดีกว่าการรักษา: Prevention is better than cure' เพราะการป้องกันจะช่วยในการลดต้นทุนการตรวจ การดูแล และการรักษา หากเพียงแต่ว่าโรคไม่ติดต่อดังกล่าวไม่สามารถระบุสาเหตุของโรคที่แท้จริงทั้งหมดได้ การกำหนดนโยบายเพื่อลดและควบคุมโรคดังกล่าวจึงเป็นไปได้ยาก

การวิจัยเกี่ยวกับโรคไม่ติดต่อดังกล่าวมีการศึกษาทั้งในเชิงลึกและเชิงกว้าง ด้านการศึกษาวิจัยทางการแพทย์ในเชิงลึก เช่นการทดลองกับกลุ่มผู้ป่วยที่ได้มีการศึกษาอย่างกว้างขวางทั้งในไทยและต่างประเทศ การศึกษาอาจมุ่งเน้นด้านการรักษาเฉพาะกลุ่มผู้เป็นโรคนั้นๆ มากกว่าการป้องกัน บางการศึกษามีการนำข้อมูลจากฐานข้อมูลคนไข้ที่เข้ามารับการรักษาในสถานพยาบาลต่างๆ มาทำการวิเคราะห์ ผู้วิจัยพบว่าข้อมูลจากกลุ่มผู้ป่วยไม่สามารถนำมาอนุมานไปสู่พฤติกรรมที่แท้จริงซึ่งเป็นสาเหตุของการเกิดโรคไม่ติดต่อดังกล่าวได้ เนื่องจากผู้ป่วยอาจมีการเปลี่ยนแปลงพฤติกรรมหลังจากทราบว่าตนเป็นโรคต่างๆ อีกทั้งปัจจัยด้านความเสี่ยงด้านพฤติกรรมของผู้ป่วยที่นำมาทำการศึกษาอาจไม่ได้มีการบันทึกไว้ และยังคงขาดข้อมูลของผู้ป่วยที่ยังไม่ได้เข้าสู่ระบบของการรักษาหรือไม่ทราบว่าตนเป็นโรคนั้นๆ ด้วยเหตุนี้สำนักงานสำรวจสุขภาพประชาชนไทย สถาบันวิจัยระบบสาธารณสุข จึงได้ทำการสำรวจและรวบรวมข้อมูล ความชุกของโรค ปัจจัยเสี่ยงทางสุขภาพ และลักษณะทางประชากรที่สำคัญ ด้วยวิธีการสัมภาษณ์ ตรวจร่างกาย และการตรวจเลือดและปัสสาวะทางห้องปฏิบัติการ เพื่อให้ได้ข้อมูลในทุกมิติ ที่เรียกว่า 'การสำรวจสุขภาพประชาชนไทยโดยการตรวจร่างกาย' ในกลุ่มประชากรทั่วไป

จากข้อมูลดังกล่าวมีการนำไปวิเคราะห์อ้างอิงเพื่อวางแผนนโยบายสาธารณสุขข้อมูลส่วนใหญ่มีการนำเสนอในรูปแบบรายงานที่ปรากฏเน้นการจำแนกปัจจัยเสี่ยงเชิงเดี่ยวหรือจำแนกแต่ละปัจจัยเสี่ยงอย่างเป็นอิสระกันในรูปตารางจำแนกความถี่แบบ 2 หรือ 3 ทาง (ดูผลการวิเคราะห์จากรายงานการสำรวจสุขภาพประชาชนไทยโดยการตรวจร่างกาย (สำนักงานสำรวจสุขภาพประชาชนไทย, 2552), (สำนักโรคไม่ติดต่อ กรมควบคุมโรค, 2556), (Aekplakorn et al., 2007)) แต่การทบทวนวรรณกรรมพบว่า ปัจจัยที่ส่งผลต่อการเกิดโรคไม่ติดต่อนั้นมีปฏิสัมพันธ์ (Interaction) ต่อกัน เช่น ถ้าทราบว่ามีผู้ที่เป็นโรคอ้วนการหาโอกาสเป็นทั้งโรคเบาหวานและไขมันในเส้นเลือดสูง หรือวิเคราะห์หาระดับปัจจัยเสี่ยงของคนที่เป็นโรคระบบหัวใจและหลอดเลือดอาจมีสาเหตุจากหลายปัจจัยได้แก่ ความดันโลหิตสูง เบาหวาน คอเลสเตอรอลรวมสูง อ้วนและการสูบบุหรี่เป็นประจำ (สำนักงานสำรวจสุขภาพประชาชนไทย, 2552) นอกจากนี้ยังพบว่าการทำงานร่วมกันระหว่างปัจจัยสามารถเป็นตัวเร่งให้เกิดโรคไม่ติดต่อต่างๆ มากขึ้น เช่น ถ้าทราบว่ามีผู้ที่เป็นโรคอ้วนและไม่ออกกำลังกาย โอกาสที่เขาจะเป็นโรคเบาหวานจะมากหรือน้อยกว่าผู้ที่เป็นโรคอ้วนและออกกำลังกาย ดังนั้นผู้วิจัยพบว่าการนำโมเดลที่สามารถช่วยวิเคราะห์ปัจจัยเสี่ยงของการเกิดโรคไม่ติดต่อโดยสามารถรวมเอาอิทธิพลการเกิดปฏิสัมพันธ์ระหว่างปัจจัยมาร่วมในการสร้างโมเดลเพื่อการวิเคราะห์หาสาเหตุการเกิดโรคไม่ติดต่อต่างๆ ที่สำคัญ โดยเฉพาะเมื่อนำมาอธิบายในรูปแบบความสัมพันธ์ด้วยกราฟจึงอาจมีความซับซ้อนและเรียกกราฟดังกล่าวว่าเครือข่าย (Network) และยิ่งเครือข่ายของโมเดลมีขนาดใหญ่เท่าไหร่ ความต้องการใช้ข้อมูลความน่าจะเป็นยิ่งมีจำนวนมากขึ้น และด้วยข้อมูลการสำรวจสุขภาพของประชาชนไทยโดยการตรวจร่ายกายจะสามารถนำมาใช้ในการสร้างโมเดลเครือข่ายแบบเบย์ เพื่อเพิ่มประสิทธิภาพและช่วยในการวิเคราะห์หาสาเหตุการเกิดโรคไม่ติดต่อจะสามารถนำไปสู่การวางแผนจัดการปัจจัยเสี่ยงเหล่านั้นต่อไปได้อย่างมีประสิทธิภาพได้ต่อไป

ด้วยเหตุนี้ผู้วิจัยจึงเสนอโมเดลเครือข่ายแบบเบย์ (Bayesian Network) ซึ่งมีลักษณะเด่นสามารถนำเอาความสัมพันธ์ระหว่างปัจจัยมานำเสนอในรูปแบบความสัมพันธ์ของเหตุและผล (Cause-effect relationship) ที่เรียกว่าในรูปแบบเครือขาย สามารถนำมาประยุกต์ใช้กับปัญหาความสัมพันธ์ระหว่างปัจจัยเสี่ยงต่างๆ ต่อการเกิดโรคไม่ติดต่อที่สำคัญ เนื่องจากขณะนี้ผู้วิจัยได้กำลังดำเนินงานวิจัยในการสร้างโมเดลเครือข่ายแบบเบย์ (Bayesian Network) จากข้อมูลการสำรวจสุขภาพอนามัยโดยการตรวจร่างกาย ซึ่งได้รับการอนุเคราะห์จากสำนักงานการสำรวจสภาวะสุขภาพของประชาชนไทย (สสท.)

สำหรับการวิจัยในครั้งนี้จะเป็นส่วนเพิ่มเติมจากงานวิจัยของกนกกรรณ์ ลี้โรจนาประภา และคณะ (Leerojanaprap, Atthirawong, Aekplakorn, & Sirikasemsuk, 2017) ที่ใช้ข้อมูลข้อมูลการสำรวจสุขภาพอนามัยโดยการตรวจร่างกาย ประกอบกับความคิดเห็นของผู้เชี่ยวชาญนำมาสร้างโมเดลเครือข่ายแบบเบย์ มาพัฒนาร่วมกับใช้เทคนิคการสร้างโครงสร้างของโมเดลด้วยวิธีการเรียนรู้ด้วยเครื่องและการกำหนดค่าความน่าจะเป็นจากข้อมูลการสำรวจสุขภาพอนามัยโดยการตรวจร่างกาย (Structure learning and Parameter learning) จากการใช้เทคนิคใหม่นี้ผู้วิจัยคาดว่าจะนำทั้ง

โครงสร้างโมเดลรวมทั้งผลการวิเคราะห์มาทำการเปรียบเทียบ นอกจากนี้ผู้วิจัยยังคาดหวังว่าผลการวิเคราะห์จากโมเดลจะช่วยวิเคราะห์หาสาเหตุหลักของการเกิดโรคไม่ติดต่อดังกล่าวอย่างเป็นระบบในบริบทของประเทศไทยซึ่งจะสามารถตอบสนองกรอบการวิจัยทางการแพทย์และสาธารณสุขในการวางแผนการป้อมปรามโรคไม่ติดต่อที่เหมาะสมกับสภาวะแวดล้อมในปัจจุบันของประทศไทย โดยมุ่งเน้นการแก้ไขที่ต้นเหตุของโรคที่แท้จริงซึ่งจะเป็นการจัดการกับปัจจัยเสี่ยงของโรคได้อย่างมีประสิทธิภาพอย่างแท้จริงในบริบทของประชาชนไทย

1.2 วัตถุประสงค์ของโครงการ

เปรียบเทียบการพยากรณ์ความชุกของโรคและสาเหตุการเกิดโรคของคนไทยในปัจจุบันโดยโมเดลเครือข่ายแบบเบย์รูปแบบต่างๆ

1.3 ขอบเขตของการวิจัย

โรคไม่ติดต่อที่นำมาศึกษาครั้งนี้เน้นเฉพาะ โรคเบาหวาน (Diabetes) โดยพิจารณาความเหมาะสมของการสร้างโมเดลจากข้อมูลการสำรวจสุขภาพของประชาชนไทยโดยการตรวจร่างกาย

กลุ่มประชากรไทย อายุ 15 – 59 ปี เท่านั้น

ข้อมูลการสำรวจสุขภาพประชาชนไทยโดยการตรวจร่างกาย มีการดำเนินการสำรวจโดยการสุ่มตัวอย่างคนไทยในช่วงอายุต่างๆ เพื่อศึกษาความชุกของปัญหาสุขภาพของประชาชนไทยโดยทำการเก็บข้อมูลด้วยวิธีการสัมภาษณ์และตรวจร่างกายจากกลุ่มตัวอย่างในช่วงอายุต่างๆ ที่อาศัยอยู่ทุกภาคของประเทศไทย การสำรวจเริ่มต้นครั้งแรกในปี พ.ศ. 2534 และดำเนินการสำรวจอย่างต่อเนื่องทุก 5 ปี การสำรวจครั้งล่าสุดคือครั้งที่ 4 ปี พ.ศ. 2551-2552[1]

1.4 ประโยชน์ที่คาดว่าจะได้รับ

การศึกษาคาดว่าจะสามารถสร้างประโยชน์ใน 2 ด้าน

1. เปรียบเทียบความแตกต่างจากการวิเคราะห์ด้วยโมเดลเครือข่ายแบบเบย์ที่สร้างขึ้นโดยวิธีการเรียนรู้ของเครื่อง (Learning Machine) และการวิเคราะห์ด้วยวิธีแบบเบย์ที่สร้างขึ้นจากความรู้ของผู้เชี่ยวชาญ (Expert Knowledge)
2. ช่วยให้ผู้บริหารด้านนโยบายสาธารณสุขสามารถวิเคราะห์หาสาเหตุหลักของปัญหาการเกิดโรคไม่ติดต่อจนนำไปสู่การกำหนดนโยบายจัดการปัจจัยเสี่ยงเพื่อเตรียมแผนป้องกัน ควบคุม ส่งเสริมสุขภาพและการรักษาอย่างมีประสิทธิภาพ

[1] ครั้งล่าสุดเป็นครั้งที่ 5 แต่ข้อมูลกำลังอยู่ระหว่างประมวลผล

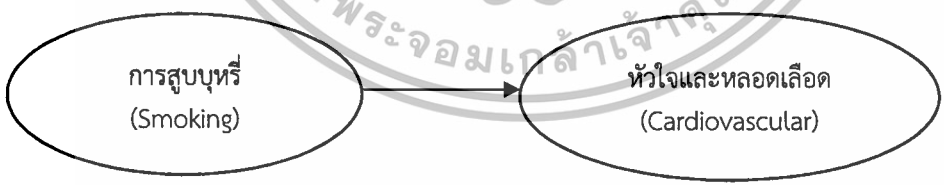
บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะนำทฤษฎีเครือข่ายเบย์มาอธิบายในมุมของการเกิดโรคเรื้อรัง เพื่อให้เห็นแนวทางการประยุกต์ใช้ของทฤษฎีต่อการเกิดโรคไม่ติดต่อที่จะนำเสนอในหัวข้อ 2-1 จนจะนำไปสู่ผลที่ได้จากการวิเคราะห์ที่จะกล่าวต่อไปในบทที่ 4 นอกจากนี้ ยังมีการนำเสนอการสร้างโมเดลเครือข่ายแบบเบย์โดยผู้เชี่ยวชาญและด้วยการเรียนรู้จากเครื่องที่สรุปไว้ในหัวข้อ 2-2 – 2-3 จากนั้นนำเสนอวิธีการสร้างต้นไม่ตัดสินใจ ในหัวข้อ 2-4 ส่วนในหัวข้อที่ 2-5 จะได้นำเสนอวิธีการตรวจสอบและเปรียบเทียบโมเดลที่ได้กับข้อมูลจริงเพื่อให้ทราบว่าโมเดลมีความสามารถในการพยากรณ์มากน้อยเพียงใด รวมทั้งทฤษฎีทางสถิติที่จะนำมาใช้เกี่ยวกับการทดสอบความเป็นอิสระในหัวข้อ 2-6 และในหัวข้อสุดท้ายเป็นการนำเสนอการทบทวนวรรณกรรมในกรณีที่มีการนำเครือข่ายแบบเบย์มาใช้วิเคราะห์ที่เกี่ยวข้อง

2.1 ทฤษฎีเครือข่ายแบบเบย์ (Bayesian Network)

2.1.1 การกำหนดโครงสร้างเครือข่ายแบบเบย์

โมเดลเครือข่ายแบบเบย์คือ โมเดลที่สามารถนำเสนอในรูปแบบความสัมพันธ์ระหว่างเหตุและผล ซึ่งมีลักษณะเช่นเดียวกับสาเหตุการเกิดโรคต่างๆ ได้ ในรูปแบบของกราฟที่ใช้ลูกศรแสดงความสัมพันธ์จากปัจจัยที่เป็นต้นเหตุสู่ปัจจัยที่เป็นผลลัพธ์ที่ทำให้เกิดโรคต่างๆ เช่น การสูบบุหรี่เป็นสาเหตุของการเกิดโรคหัวใจและหลอดเลือด (Cardiovascular) สามารถแสดงได้ด้วยกราฟ ดังรูป 2-1



รูป 2-1 ตัวอย่างโครงสร้างโมเดลเครือข่ายแบบเบย์ (อย่างง่าย)

โดยทั่วไปโครงสร้างเครือข่ายแบบเบย์สามารถแบ่งกลุ่มตัวแปรเป็น 3 กลุ่มคือ

1. ปัจจัยที่เป็นต้นเหตุของปัญหา (Root cause) ซึ่งเป็นตัวแปรที่ไม่มีตัวแปรอื่นอยู่ก่อนหน้า
2. ปัจจัยที่เป็นสาเหตุย่อย (Effect variable as Child variable) เป็นตัวแปรที่มีลูกศรจากตัวแปรอื่นพุ่งเข้าและมีลูกศรพุ่งออกไปยังตัวแปรที่เป็นผลที่เกิดจากตัวแปรดังกล่าว
3. ปัจจัยผลกระทบ (Effect) คือปัจจัยที่เป็นตัวแปรที่เป็นตัวแปรสุดท้ายที่ไม่มีลูกศรพุ่งออก

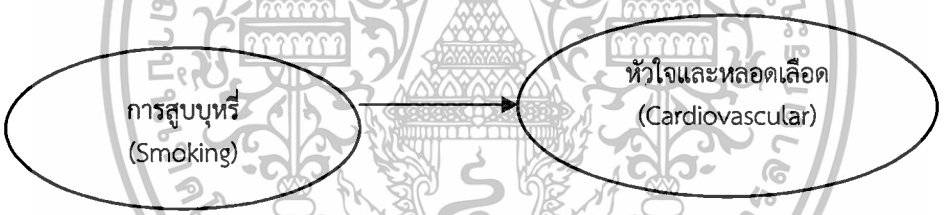
2.1.2 การกำหนดค่าความไม่แน่นอน – การกำหนดค่า โมเดลเครือข่ายแบบเบย์

ความน่าจะเป็น (Probability) คือ โอกาสของแต่ละสถานะ (State) ของตัวแปร ซึ่งบ่งบอกถึงระดับของความเป็นไปได้ที่มีค่าระหว่าง 0 ถึง 1 เพื่อแสดงความน่าจะเป็นของการเกิดขึ้นในแต่ละสถานะ (State) สำหรับแต่ละตัวแปร

"การสูบบุหรี่ (Smoking)" เป็นตัวแปรต้นเหตุของปัญหา (Root cause) ของ "การเกิดโรคหัวใจและหลอดเลือด (Cardiovascular)" ที่เป็นตัวแปรผลลัพธ์ (Output) โดยที่ทราบว่ามีเพียงบางส่วนของประชาการที่สูบบุหรี่ ซึ่งสมมุติว่ามีร้อยละ 19.9 ของประชากร แทนด้วย $P(X=x)=P(X=smoking)$ แสดงดังตารางซ้ายมือในรูป 2-5

นอกจากนี้ยังสามารถแสดงผลกระทบที่ไม่แน่นอนของความสัมพันธ์ระหว่างปัจจัยได้ด้วย เช่น ไม่จำเป็นที่คนสูบบุหรี่ทุกคนจะเป็นโรคหัวใจและหลอดเลือด ด้วยเหตุนี้ระดับความสัมพันธ์ระหว่างปัจจัยทั้งสองสามารถกำหนดได้ด้วยระดับความน่าจะเป็นหรือสัดส่วนของผู้ที่สูบบุหรี่และเป็นโรคหัวใจและหลอดเลือดต่อจำนวนผู้สูบบุหรี่ทั้งหมด จากรูปแบบผลลัพธ์ที่ไม่แน่นนอน (Uncertain effect) สามารถแสดงได้ด้วยตารางความน่าจะเป็นแบบมีเงื่อนไข (Conditional probability) ซึ่งเป็นตารางทางขวามือ (รูป 2-2) เช่นพบว่า ร้อยละ 34.1 ของผู้สูบบุหรี่เป็นโรคหัวใจและหลอดเลือดแทนด้วย

$$P(Y=y|X=x)=P(Y=\text{positive}\,|\,X=\text{smoking})=0.341$$



การสูบบุหรี่ (Smoking)	ความน่าจะเป็น
สูบ (Smoking)	0.199
ไม่สูบ (Nonsmoking)	0.801

หัวใจและหลอดเลือด (Cardiovascular)	การสูบบุหรี่ (Smoking)	
	สูบ	ไม่สูบ
เป็นโรค (Positive)	0.341	0.079
ไม่เป็นโรค (Negative)	0.659	0.921

หมายเหตุ : ผลรวมแนวคอลัมน์เป็น 1

รูป 2-2 ตัวอย่างการกำหนดค่าความน่าจะเป็นของโมเดลเครือข่ายแบบเบย์ ด้วยตารางความน่าจะเป็น (Probability Table : PT) และตารางความน่าจะเป็นแบบมีเงื่อนไข (Conditional Probability Table : CPT)

โดยทั่วไปแล้วตัวแปรสามารถอธิบายได้จากตัวแปรต้นเหตุของปัญหา (Root cause) และตัวแปรผลลัพธ์ (Effect) ซึ่งสามารถวัดได้จากตารางความน่าจะเป็น (Probability Table : PT) และตารางความน่าจะเป็นแบบมีเงื่อนไข (Conditional Probability Table : CPT)

ถ้าเป็นตัวแปรต้นทางจะไม่มีลูกศรจากตัวแปรอื่นเข้าไปในตัวมันดังนั้นจึงเป็นอิสระจากเงื่อนไข ดังนั้นจึงไม่มีเงื่อนไขกับตัวแปรอื่น ๆ ความน่าจะเป็นของต้นเหตุของปัญหา อยู่ในสถานะ x, สามารถกำหนดตัวเลข ในตารางความน่าจะเป็น (PT)

ถ้าเป็นตัวแปรผลเป็นตัวแปรย่อย (Child variable) เนื่องจากมักมีตัวแปรหลักที่เชื่อมโยงอยู่ (Parent variable) จะเป็นตัวแปรตามตัวแปรอื่น ๆ ดังนั้นความน่าจะเป็นของการเกิดตัวแปรผลขึ้นอยู่กับสาเหตุของมัน ความน่าจะเป็นของตัวแปรผลกระทบที่อยู่ในสถานะที่กำหนดสามารถกำหนดโดยสถานะที่รู้จักกันในตัวแปรหลักซึ่งอาจจะมีมากว่า 1 ตัวและเป็นตัวเลขในตารางความน่าจะเป็นเงื่อนไข (CPT) นอกจากนี้ยังมี 2 กรณีพิเศษ คือถ้าเหตุการณ์สาเหตุเกิดขึ้นจะมีผลต่อตัวแปรผลกระทบในลักษณะที่จะเกิดขึ้นอย่างแน่นอน (ความน่าจะเป็น = 1) หรือไม่เกิดขึ้นแน่นอน (ความน่าจะเป็น = 0) นี้เรียกว่าปัจจัยความเชื่อมั่นซึ่งสามารถแสดงถึงความสัมพันธ์เชิง Deterministic แบบสมบูรณ์

2.1.3 การอนุมานด้วยโมเดลเครือข่ายแบบเบย์

2.1.3.1 หลักการเบื้องต้นของการอนุมานด้วยเครือข่ายแบบเบย์

เมื่อตัวเลขความน่าจะเป็นถูกนำเข้าสู่โมเดลเรียบร้อยแล้ว ขั้นต่อไปคือการใช้ประโยชน์จากโมเดลโดยการวิเคราะห์ผลจากทฤษฎีเบย์ที่เป็นทฤษฎีพื้นฐานของโมเดลเครือข่ายแบบเบย์

สมมุติให้ ตัวแปร A เป็นตัวแปรสาเหตุเพียงตัวแปรเดียวของตัวแปร B เช่น จากตัวอย่างข้างต้น A คือ ‘การสูบบุหรี่’ และ B คือ ‘โรคหัวใจและหลอดเลือด’ การคำนวณโดยใช้ทฤษฎีเบย์ ทำได้โดยสมการ (2-1)

$$P(A|B)=\frac{P(B|A)*P(A)}{P(B)} \qquad (2\text{-}1)$$

เมื่อ

$P(A)$ คือ ความน่าจะเป็นของการเกิดเหตุการณ์ *A*

$P(B|A)$ คือ ความน่าจะเป็นแบบมีเงื่อนไขของเหตุการณ์ *B* โดยกำหนดว่าเหตุการณ์ *A* เกิดขึ้น

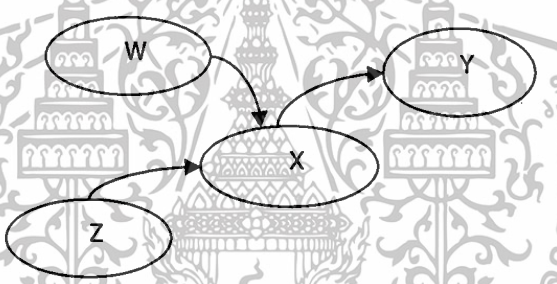
รายละเอียดการคำนวณและการอนุมานจากทฤษฎีสามารถอ่านได้เพิ่มเติมจาก *Jensen & Nielsen* (2007)

2.1.4 โมเดลเครือข่ายแบบเบย์เมื่อมีความซับซ้อนมากยิ่งขึ้น

"ความน่าจะเป็นในปัจจุบัน" (Current probability) vs. "ความน่าจะเป็นที่เปลี่ยนไปของแต่ละปัจจัยเมื่อทราบสถานะของตัวแปรเป้าหมาย" (Adjusted probability)

โดยทั่วไปแล้วโมเดลเครือข่ายแบบเบย์จะมีความซับซ้อนมากขึ้นเมื่อเป็นโครงสร้างของตัวแปรหลายตัวและทำให้มีลูกศรจำนวนมากเชื่อมโยงในโมเดล ซึ่งทำให้การอนุมานโดยใช้โมเดลเครือข่ายแบบเบย์เป็นทำได้ยากขึ้น โดยเฉพาะอย่างยิ่งในขั้นตอนการหาการแจกแจงความเป็นไปได้ร่วม (Joint probability distribution) เพื่อนำมาคำนวณค่าความน่าจะเป็นของแต่ละตัวแปรที่จะเกิดขึ้น (Marginal probability) และความน่าจะเป็นส่วนหลัง (Marginal posterior probability) ตามทฤษฎีเครือข่ายแบบเบย์ที่นำเอาทฤษฎีความเป็นอิสระตามเงื่อนไข (Conditional independence) มาใช้ในการอนุมานในโมเดลเครือข่ายแบบเบย์ โดยสามารถแสดงการอธิบายในรายละเอียดในหัวข้อนี้

ตัวอย่าง การอนุมานจากโครงสร้าง โมเดล BN จาก 4 ตัวแปร แสดงดัง รูป 2-3.



รูป 2-3 ตัวอย่างโมเดลเครือข่ายแบบเบย์ ที่ประกอบด้วย 4 ตัวแปร

โดยทั่วไปเครือข่ายแบบเบย์ ที่ประกอบด้วย n ตัวแปร ที่เชื่อมโยงกันเป็นลำดับด้วยลูกศรทางเดียวจนเกิดเป็นเครือข่าย (Network) แสดงด้วย กราฟระบุทิศทาง (Directed graphs) แสดงได้โดย $X_i \longrightarrow X_{i+1}$ for *i = 1, 2, 3, ... , n-1* โดยที่ไม่ทำให้เกิดรูปแบบโครงสร้างแบบไซเคิล (Cycle) ซึ่งจะเรียกว่า "a Directed Acyclic Graph (DAG)" ดังนั้นจากกฎ Chain rule หรือ Markov property ทำให้สามารถคำนวณค่า การแจกแจงความเป็นไปได้ร่วม (Joint probability distribution) ของตัวแปร n ตัวที่เกิดจากผลคูณของความน่าจะเป็นแบบมีเงื่อนไขที่ถูกกำหนดในเครือข่ายแบบเบย์ โดยมีรูปทั่วไป แสดงได้ดังสมการ (2-2)

$$P(X_1,...,X_n) = \prod_{i=1}^{n} P(X_i|X_1,...,X_{i-1}) \qquad (2\text{-}2)$$

จากตัวอย่างโครงสร้างโมเดลเครือข่ายแบบเบย์ที่กำหนดในรูป 2-3 ซึ่งจากโครงสร้างดังกล่าวพบว่า ตัวแปร Y เป็นตัวแปรที่ไม่มีลูกศรออกจากตัวแปรดังกล่าว ซึ่งเรียกว่า "Top

variable" และรูปแบบโครงสร้างของโมเดล ที่เกิดจากความสัมพันธ์ของตัวแปร *W, X , Z* ที่ไม่ทำให้เกิดโครงสร้างแบบ ไซเคิล (Cycle) ดังนั้นด้วยกำของ Chain rule ทำให้สามารถคำนวณค่าความน่าจะเป็นร่วม (Joint probability) ของตัวแปร *W, X , Y, Z* ได้ดังนี้

$$\begin{aligned} P(W,X,Y,Z) &= P(Y|W,X,Z)P(X|W,Z)P(W|Z)P(Z) \\ &= P(Y|W,X,Z)P(X|W,Z)P(W)P(Z) \end{aligned} \qquad \text{(2-3)}$$

จะเห็นได้ว่าตัวแปรที่เป็นเงื่อนไขเกิดจากการรวมเอาตัวแปรที่เป็นตัวแปร Parent variables ของตัวแปรนั้นๆ กับตัวแปร Non-descendent variables เท่านั้น ซึ่งสามารถพิจารณาได้จากสมการโครงสร้างของกราฟที่สร้างขึ้น

ยกตัวอย่างเช่น การกำหนดพจน์แรกของความน่าจะเป็นแบบมีเงื่อนไขจาก Top variable: Y โดยใช้สัญลักษณ์ $P(Y|W,X,Z)$ เนื่องจากตัวแปร *W, X , Z* ทั้ง 3 ตัวเป็น Non-descendent variables โดยจากโครงสร้างจะเห็นได้ว่าเฉพาะตัวแปร *X* ที่เป็น Parent ของ ตัวแปร *Y* โดยที่ *W, Z* เป็น Non-descendent variables

ดังนั้นจากเงื่อนไขที่กำหนดในสมการ (2-2) สามารถเขียนในรูปใหม่โดยแบ่งเงื่อนไขเป็น 2 แสดงดังสมการ (2-4)

$$P(X_1,...,X_n) = \left[\prod_{i=1}^{n} P(X_i \,|\, \mathbf{PA}(X_i), \mathbf{Non\_Descendent}(X_i))\right] \qquad \text{(2-4)}$$

เมื่อ

$P(X_1,...,X_n)$ คือ ความน่าจะเป็นร่วม (Joint probability distribution),

$\mathbf{PA}(X_i)$ คือ เซทของตัวแปร parent variables ของตัวแปร X_i,

$\mathbf{Non\_Descendent}(X_i)$ คือ เซทของตัวแปรทั้งหมด ใน DAG ที่ไม่ใช่ตัวแปรที่ไม่ใช่ $\mathbf{PA}(X_i)$ และ $\mathbf{Descendent}(X_i)$ ของตัวแปร X_i

ขั้นต่อไปเป็นการนำเสนอคุณสมบัติของโมเดลเครือข่ายแบบเบย์ที่สำคัญที่เรียกว่า "Conditional independence" ซึ่งเป็นการอธิบายต่อเนื่องจากสมการ (2-4) ที่กำหนดให้ X_i เป็น Conditionally independent จากเซทของตัวแปร $\mathbf{Non\_Descendent}(X_i)$ โดยที่ทราบว่า เซท $\mathbf{PA}(X_i)$ คือ parents ของตัวแปร X_i ดังนั้นสมการ (2-6) สามารถลดรูปความซับซ้อนของกลุ่มตัวแปรเงื่อนไขลง โดยพิจารณาเฉพาะ $\mathbf{PA}(X_i)$ เป็นเงื่อนไขเฉพาะตัวแปร parents ของตัวแปร X_i ทำให้สมการลดรูปลงเหลือสมการ (2-5)

$$P(X_1,...,X_n)=\prod_{i=1}^{n}P(X_i|\mathbf{PA}(X_i)) \qquad (2\text{-}5)$$

จากตัวอย่างจากโครงสร้างโมเดลที่กำหนดสามารถหาค่าความน่าจะเป็นร่วม (Joint probability) โดยอาศัย "Conditional independence" ที่ใช้กำหนดความน่าจะเป็นร่วมดังสมการ (2-5) ได้ดังสมการ (2.6)

$$P(W,X,Y,Z)=P(Y|X)P(X|W,Z)P(W)P(Z) \qquad (2\text{-}6)$$

ลำดับถัดไปคือการคำนวณค่าความน่าจะเป็นของแต่ละเหตุการณ์ที่สามารถคำนวณได้จากการแจกแจงความน่าจะเป็นร่วม (Joint probability distribution) ผ่านสมการ (2-6) ซึ่งการคำนวณความน่าจะเป็นร่วม (Joint probability) สามารถคำนวณได้จากความน่าจะเป็นของแต่ละตัวแปรที่จะเกิดขึ้น (Marginal probability)

ตัวอย่าง ความน่าจะเป็นภายหลังของแต่ละตัวแปรที่จะเกิดขึ้น (Marginal posterior probability) ของตัวแปร X โดยกำหนด ค่าของตัวแปร Y สามารถสังเกตได้ ดังนั้น $P(X|Y)$, กำหนดดังสมการ (2-4) as:

$$P(X|Y)=\frac{P(X,Y)}{P(Y)}$$

เมื่อ

$$P(X,Y)=\sum_{W}\sum_{Z}P(W,X,Y,Z)$$

$$P(Y)=\sum_{X}\sum_{W}\sum_{Z}P(W,X,Y,Z)$$

จากตัวอย่างที่กำหนดโดยตัวแปร 4 ตัวแปร อาจต้องอาศัยกระบวนการคำนวณที่ค่อนข้างมาก จึงไม่นำมาแสดงการคำนวณในที่นี้ ซึ่งผู้อ่านสามารถดูรายละเอียดได้ดังหนังสือต่างๆ เช่น (Kjaerulff & Madsen, 2008; Lauritzen & Spiegelhalter, 1988) ด้วยปัญหาการคำนวณที่ยุ่งยากซับซ้อนการอนุมานด้วยโมเดลเครือข่ายแบบเบย์ จึงมีการคำนวณด้วย Algorithm และซอฟแวร์ต่างๆ ที่สร้างขึ้นเพื่อการศึกษาหรือเพื่อการค้า

2.2 การสร้างโมเดลเครือข่ายแบบเบย์โดยผู้เชี่ยวชาญ

การสร้างโมเดลเครือข่ายแบบเบย์สามารถนำเสนอในรูปแบบเหตุและผลที่สามารถนำความรู้และความเชี่ยวชาญมาถ่ายทอดในรูปแบบกราฟ ที่เรียกว่า Mental model จาก Expert domain มีการศึกษาที่มีการนำเสนอประบวนการมนการสร้างเครือข่ายแบบเบย์จากผู้เชี่ยวชาญไว้หลายงาน เช่น Kudikyala, Bugudapu, & Jakkula (2018) ใช้เทคนิคนี้ใช้วิธีในการจับภาพความสัมพันธ์เชิงสาเหตุ / อิทธิพลระหว่างโหนดความน่าจะเป็นจากผู้เชี่ยวชาญ จากนั้นจะสร้างกราฟกำกับโดยใช้อัลกอริธึม Pathfinder อาจมีการสร้างเครือข่าย Pathfinder แบบเอกฉันท์หากมีผู้เชี่ยวชาญหลายคน

2.3 การสร้างโมเดลเครือข่ายแบบเบย์ด้วยการเรียนรู้จากโครงสร้าง

เครือข่ายแบบเบย์นอกจากจะใช้ความเห็นของผู้เชี่ยวชาญในการกำหนดโครงสร้างของโมเดลเครือข่ายแบบเบย์แล้ว ยังสามารถใช้การเรียนรู้จากเครื่องเพื่อกำหนดโครงสร้างและประมาณค่าความน่าจะเป็นสำหรับสร้างตัวแบบได้อีกด้วย

กระบวนการในการสร้างโมเดลเครือข่ายแบบเบย์แบบเรียนรู้ด้วยเครื่องนั้นจากการทบทวนวรรณกรรม (Duijm, 2009) มีการดำเนินการดังต่อไปนี้

(i) Searching for the optimal structure, i.e., a directed acyclic graph that most adequately fits the learning data or the process under study;

(ii) Computing the values of conditional probability tables of the BN for the corresponding nodes of this graph.

Algorithm ในการสร้างโมเดลเครือข่ายแบบเบย์มีการศึกษาอย่างอย่างกว้างขวางในแต่ Algorithm ค่อนข้างซับซ้อนและมักต้องใช้โปรแกรมช่วยในการสร้างโมเดลโดยเฉพาะโมเดลที่มีการเรียนรู้จากข้อมูลจำนวนมากซึ่งส่วนใหญ่จะเป็นโปรแกรมที่มีราคาแพง หรือหากเป็นโปรแกรมสำหรับการทดลองใช้ก็จะมีข้อจำกัดทั้งจำนวนข้อมูลที่ใช้และ และจำนวนตัวแปรที่จะนำมาสร้างโมเดล ด้วยข้อจำกัดด้านงบประมาณผู้วิจัยจำเป็นต้องเลือกใช้โปรแกรมหลายโปรแกรมเพื่อให้ได้ผลลัพธ์ที่ต้องการตามวัตถุประสงค์ของงานวิจัยนี้

2.4 การสร้างต้นไม้ตัดสินใจ (Decision Tree)

ต้นไม้ตัดสินใจเป็นเทคนิคหนึ่งในทฤษฎีการตัดสินใจ ที่สามารถใช้ได้กับข้อมูลที่มีลักษณะเป็นเชิงกลุ่ม และนำมาประยุกต์ใช้กับกฎแบบ ถ้า-แล้ว (If-then) เพื่อให้สามารถอ่านแล้วเข้าใจการตัดสินใจของต้นไม้ได้ตามขั้นตอนการคิดที่ชัดเจน เข้าใจง่ายในรูปแบบของแผนผังต้นไม้

ในการเรียนรู้ของเครื่อง (Machine learning) จากต้นไม้ตัดสินใจ เป็นโมเดลทางคณิตศาสตร์ที่ใช้ทำนายประเภทของวัตถุโดยพิจารณาจากลักษณะของวัตถุ โหนดภายใน (Inner node) ของต้นไม้จะแสดงตัวแปร ส่วนกิ่งจะแสดงค่าที่เป็นไปได้ของตัวแปร ส่วนโหนดใบจะแสดงประเภทของวัตถุ

ต้นไม้การตัดสินใจจะทำการจัดกลุ่ม (Classify) ชุดข้อมูลนำเข้าในแต่ละกรณี แต่ละโหนด (Node) ของต้นไม้การตัดสินใจคือตัวแปร (Attribute) ต่างๆ ของชุดข้อมูล เช่นหากต้องการตัดสินใจว่าเขาจะป่วยเป็นโรคหรือไม่ (ตัวแปรตาม) ก็จะมีตัวแปรต้นที่จะต้องพิจารณาคือ ปัจจัยที่ส่งผลก่อให้เกิดโรค เช่น อายุ ปัจจัยพันธุกรรม เป็นต้น ซึ่งแต่ละตัวแปรนั้นก็จะมีค่าของตัวเอง (Value) เกิดเป็นชุดของตัวแปร-ค่าของตัวแปร (Attribute-value pair) การทำนายประเภทด้วยต้นไม้ตัดสินใจ จะเริ่มจากโหนดราก แล้วจึงตามกิ่งของต้นไม้ที่กำหนดค่า เพื่อไปยังโหนดลูกถัดไป การทดสอบนี้จะกระทำไปจนกระทั่งเจอโหนดใบซึ่งจะแสดงผลการทำนาย

2.5 การตรวจสอบความสามารถในการพยากรณ์ของตัวแบบ

ในการพยากรณ์การเกิดโรคด้วยโมเดลต่างๆ ต้องตรวจสอบความถูกต้องในการพยากรณ์และค่าที่แท้จริงโดยเปรียบเทียบจากค่าสถิติดังต่อไปนี้

2.5.1 เส้นโค้ง ROC และ AUC

ความถูกต้องของการตัดสินใจเป็นปัจจัยหลักในการตัดสินใจเลือกรูปแบบการตัดสินใจที่เหมาะสม ความสามารถในการทำนายสามารถวัดด้วยโดยการประเมินจากเส้นโค้ง ROC โดยเส้นโค้ง ROC แสดงถึงความสัมพันธ์ระหว่างเปอร์เซ็นต์ความถูกต้องเชิงบวก (True positive) และเปอร์เซ็นต์ของความผิดพลาดเชิงบวก (Fault positive) นอกจากนี้ยังใช้พื้นที่ใต้เส้นโค้ง ROC ที่เรียกว่า AUC โดยใช้เพื่อแสดงประสิทธิภาพโดยรวมในการจำแนกประเภท ค่า AUC อยู่ระหว่าง 0 ถึง 1 ถ้าค่า AUC เท่ากับ 1 แสดงว่ามีประสิทธิภาพที่ดีโดยไม่มีข้อผิดพลาด ในทางกลับกันเมื่อ AUC มีขนาดเล็กมากและมีค่าใกล้เคียงกับ 0 ตัวแบบมักมีการคาดการณ์ที่ไม่ถูกต้อง เมื่อค่า AUC ใกล้เคียง 0.5 โมเดลสามารถทำการสุ่ม (Marcot, 2012)

2.5.2 คอนฟูชั่น แมททริก (Confusion matrix)

คอนฟูชั่น แมททริกเป็นตารางที่มักใช้ในการอธิบายถึงประสิทธิภาพของแบบจำแนก (Classifier) ในชุดของข้อมูลที่ทราบค่าจริง คอนฟูชั่น แมททริกแบบง่ายๆ ที่ใช้เพื่อจำแนกกลุ่ม 2 กลุ่ม แสดงดังตารางที่ 2-1

ตาราง 2-1 คอนฟูชั่น แมททริก (Confusion matrix) ในการจำแนก 2 กลุ่ม

ค่าแท้จริง (Actual)	**ค่าพยากรณ์ (Predictive)**		**เกณฑ์การพิจารณา**	
	Positive	**Negative**		
Positive			TPR = ความไว (Sensitivity) = ความระลึก (Recall) $TPR = \frac{TP}{TP + FN}$	$FNR = \frac{FN}{TP + FN}$
Negative			ความจำเพาะ (Specification) TNR = TN / (FP + TN)	$FPR = \frac{FP}{FP + TN}$
เกณฑ์การพิจารณา	Precision $P = \frac{TP}{TP + FP}$	ค่าทำนายเมื่อผลเป็นลบ (NPV) = TN / (FN + TN)	Accuracy = (TP+TN)/(TP+FN+FP+TN)	

เมื่อ True Positives (TP): จำนวนข้อมูลที่จำแนกถูกว่าเป็นโรคและเขาเป็นโรคจริง

True Negatives (TN): จำนวนข้อมูลที่จำแนกถูกว่าไม่เป็นโรคและเขาไม่เป็นโรคจริง

False Positives (FP): จำนวนข้อมูลที่จำแนกผิดว่าเป็นโรคแต่จริงๆ แล้วเขาไม่เป็นโรคจริง

False Negatives (FN): จำนวนข้อมูลที่จำแนกผิดว่าไม่เป็นโรคแต่จริงๆ แล้วเขาเป็นโรคจริง

จากตารางคอนฟูชั่น แมททริก สามารถนำมาคำนวณค่าเพื่อตรวจสอบความสามารถในการพยากรณ์เพื่อการจำแนกกลุ่มได้กลายเกณฑ์ ซึ่งในงานวิจัยนี้

1. อัตราความถูกต้องเชิงบวก (*True Positive Rate : TPR)*

$$TPR = \frac{TP}{TP + FN} \qquad (2\text{-}7)$$

TPR = <u>จำนวนข้อมูลที่พยากรณ์ถูกว่าอยู่กลุ่มที่สนใจ</u>
จำนวนข้อมูลที่แท้จริงในกลุ่มที่สนใจ

2. อัตราความถูกต้องเชิงลบ (*True Negative Rate : TNR)*

$$TNR = \frac{TN}{FP + TN} \quad (2\text{-}8)$$

TNR = จำนวนข้อมูลที่พยากรณ์ถูกว่าอยู่กลุ่มที่ไม่สนใจ / จำนวนข้อมูลที่แท้จริงในกลุ่มที่ไม่สนใจ

3. อัตราความผิดพลาดเชิงบวก (*Fault Positive Rate : FPR)*

$$FPR = \frac{FP}{FP + TN} \quad (2\text{-}9)$$

FPR = จำนวนข้อมูลที่พยากรณ์ผิดว่าอยู่กลุ่มที่สนใจ / จำนวนข้อมูลที่แท้จริงในกลุ่มที่ไม่สนใจ

4. อัตราความผิดพลาดเชิงลบ (*Fault Negative Rate : FNR)*

$$FNR = \frac{FN}{TP + FN} \quad (2\text{-}10)$$

FNR = จำนวนข้อมูลที่พยากรณ์ผิดว่าอยู่กลุ่มที่ไม่สนใจ / จำนวนข้อมูลที่แท้จริงในกลุ่มที่สนใจ

5. ค่าความแม่นยำ (*Precision*)

$$Presition = \frac{TP}{TP + FP} \quad (2\text{-}11)$$

Precision = จำนวนข้อมูลที่พยากรณ์ถูกว่าอยู่กลุ่มที่สนใจ / จำนวนข้อมูลที่ทำนายว่าอยู่ในกลุ่มที่สนใจ

6. ความถูกต้อง (*Accuracy*)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2\text{-}12)$$

จำนวนข้อมูลที่ทั้งหมด

7. *ความไว* (Sensitivity) *หรือ ความระลึก (Recall)*

$$Sensitivity = \frac{TP}{TP + FN} = TPR \quad (2\text{-}13)$$

ความไวและความจำเพาะเป็นค่าวัดทางสถิติที่ใช้ประเมินประสิทธิภาพของการทดสอบที่ให้ผลเป็นสองส่วน (เช่น เป็นบวกและลบ) โดย

ความไว (Sensitivity) คือ สัดส่วนของผลบวกที่เป็นจริงสำหรับภาวะนั้น ๆ (เช่น สัดส่วนของการตรวจพบโรคในผู้ที่ป่วยจริง) มีคำอื่น ๆ รวมทั้ง อัตราผลบวกจริง (TPR: True positive rate), recall, probability of detection ซึ่งใช้ในสาขาต่าง ๆ

8. *ความจำเพาะ (Specificity)*

$$Specification = \frac{TN}{FP + TN} = TNR \quad (2\text{-}14)$$

ความจำเพาะ (Specificity) คือ สัดส่วนของผลลบที่เป็นจริงสำหรับภาวะนั้น ๆ (เช่น สัดส่วนของการตรวจไม่พบโรคในผู้ที่ไม่ป่วย) มีคำอื่น ๆ รวมทั้ง อัตราผลลบจริง (TNR: True negative rate)

ความไวจึงมีประโยชน์ในการวินิจฉัยแยกกันความผิดพลาดเชิงลบ (False negative) เพราะว่าการทดสอบยิ่งไวเท่าไร โอกาสการได้ผลลบ (เช่น การพบว่าไม่มีโรค) ที่ไม่เป็นจริง (เช่น บุคคลจริง ๆ มีโรค) ก็น้อยลงเท่านั้น และดังนั้น ถ้าความไวอยู่ที่ 100% โอกาสได้ผลความผิดพลาดเชิงลบ ก็อยู่ที่ 0% และความจำเพาะจึงมีประโยชน์ในการยืนยันภาวะที่มี โดยป้องกันการเกิดอัตราความผิดพลาดเชิงบวก (False positive) เพราะว่าการทดสอบยิ่งจำเพาะเท่าไร โอกาสการได้ผลบวก (เช่น การพบว่ามีโรค) ที่ไม่เป็นจริง (เช่น บุคคลจริง ๆ ไม่มีโรค) ก็น้อยลงเท่านั้น และดังนั้น ถ้าความจำเพาะอยู่ที่ 100% โอกาสได้ผลบวกปลอมก็อยู่ที่ 0%

2.5.3 ค่า F1

F1 คำนวณจากค่าเฉลี่ย ฮาร์โมนิคระหว่างค่าความแม่นยำ (Precision) และความระลึก (Recall) ใช้เพื่อเปรียบเทียบความสามารถในการแบ่งกลุ่ม แสดงได้ดังสูตรนี้

$$F1 = 2\left(\frac{\Pr ecision * \mathrm{Re}\, call}{\Pr ecision + \mathrm{Re}\, call}\right) \qquad (2\text{-}15)$$

2.5.4 ค่า G-Mean

การคำนวณค่าเฉลี่ยเลขา G-mean คือค่าเฉลี่ยเรขาคณิตระหว่างค่าความไว (Sensitivity) และความจำเพาะ (Specificity) แสดงได้ดังสูตรนี้

$$G - Mean = \sqrt{Sensitivity \times Specificity} \qquad (2\text{-}16)$$

2.6 การทดสอบความเป็นอิสระ (Test of Independence)

การใช้สถิติทดสอบ Pearson's chi-square test โดยใช้ χ^2 statistic เป็นสถิติไม่ใช้พารามิเตอร์และใช้เพื่อทดสอบความเป็นอิสระระหว่างตัวแปรเชิงกลุ่ม 2 ตัวแปร ความสัมพันธ์อย่างมีนัยสำคัญระหว่างตัวแปร 2 ตัวแปรที่เป็นตัวแปรเหตุและผลที่ทำให้เกิดโรคเบาหวาน สามารถกำหนดได้ด้วยการทดสอบนี้โดยกำหนดสมมุติฐานว่าง แสดงว่าตัวแปรทั้งสองไม่ขึ้นอยู่ต่อกัน ค่าสถิติทดสอบ χ^2 สามารถคำนวณได้ดังนี้:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \qquad (2\text{-}17)$$

$$E_{ij} = \frac{(r_i)(c_j)}{n} \qquad (2\text{-}18)$$

เมื่อ

O_{ij} = ความถี่ที่สังเกตได้ แถวที่ i^{th} คอลัมน์ j^{th}

E_{ij} = ความถี่คาดหวัง แถวที่ i^{th} คอลัมน์ j^{th}

r_i = ผลรวมความถี่แถวที่ i^{th}

c_j = ผลรวมความถี่คอลัมน์ที่ j^{th}

n = ผลรวมทั้งหมด

r = จำนวนแถว

c = จำนวนคอลัมน์

ค่า χ^2 ที่คำนวณได้ มีค่าองศาความเป็นอิสระ (r-1)(n-1) สามารถนำมาใช้เพื่อคำนวณค่า p-value และนำค่าที่ได้มาเปรียบเทียบกับ α หากค่า p-value < α สามารถสรุปได้ว่า ตัวแปรทั้งคู่ไม่เป็นอิสระกัน นอกจากนี้การคำนวณค่าสถิติทดสอบ χ^2 มีข้อจำกัดที่จำนวนเซลล์ที่มี $E_{ij} < 5$ ไม่ควรเกิน20% ของจำนวนเซลล์ที่มีทั้งหมด ซึ่งในที่นี้ใช้โปรแกรม SPSS version 19.0 for Windows ในการคำนวณ

2.7 วรรณกรรมที่เกี่ยวข้อง (Literature Review)

ผลการสำรวจสุขภาพประชาชนไทยโดยการตรวจร่างกายครั้งที่ 1- 4 ได้มีการจัดทำในรูปแบบรายงาน ซึ่งสามารถ download ได้จากเว็บไซท์ของ 'สำนักงานพัฒนาระบบข้อมูลข่าวสารสุขภาพ' ข้อมูลจากการสำรวจยังได้นำไปใช้โดยหน่วยงานที่เกี่ยวข้องเพื่อรายงานการเฝ้าระวังโรค (อมรา ทองหงษ์, กมลชนก เทพสิทธา, & ภาคภูมิ จงพิริยะอนันต์, 2556) หรือรายงานประจำปี (สำนักโรคไม่ติดต่อ กรมควบคุมโรค, 2556) รวมทั้งนำเสนอในรูปบทความทางวิชาการระดับนานาชาติ (ยกตัวอย่างเช่น Aekplakorn et al., 2007; Danaei et al., 2011)

จากการทบทวนวรรณกรรมยังพบว่าวิธีเครือข่ายแบบเบย์เคยนำมาใช้เพื่อวินิจฉัยหรือพยากรณ์การเกิดโรคต่างๆ จากฐานข้อมูลผู้ป่วยที่เข้ารับการรักษาในโรงพยาบาล เช่น การพยากรณ์ความเสี่ยงการเป็นโรคมะเร็งเต้านม (Burnside et al., 2006) การพยากรณ์การเกิดโรคโลหิตจาง (Sebastiani et al., 2007) การพยากรณ์การเกิดโรคเบาหวานประเภทที่ 2 (Yang Guo, Guohua Bai, & Yan Hu, 2012) หรือการประเมินความเสี่ยงต่อสุขภาพจากมลพิษทางอากาศ (Liu, Lu, Chen, & Shen, 2011) ผู้วิจัยพบว่าการศึกษาภายใต้ฐานข้อมูลผู้ป่วยที่มีการบันทึกตามสถานพยาบาลต่างๆ มีข้อจำกัดเพราะไม่สามารถอนุมานไปสู่ความชุกระดับประชากรผู้ที่เป็นโรคแต่ยังไม่ทราบหรือไม่ได้เข้าสู่ระบบของการรักษา คนกลุ่มนี้เป็นกลุ่มที่มีความสำคัญเพราะพวกเขายังไม่มีการปรับพฤติกรรมเสี่ยงเนื่องจากยังไม่ทราบว่าตนเองเป็นโรค

เมื่อมุ่งทบทวนวรรณกรรมเฉพาะที่นำฐานข้อมูลระดับประชากรที่ได้จากการสุ่มตัวอย่างมาจากประชากรระดับประเทศเพื่อมาสร้างโมเดลเครือข่ายแบบเบย์พบว่า โมเดลดังกล่าวไม่รวมปัจจัยความเสี่ยงจากความแตกต่างด้านลักษณะประชากรที่สำคัญ เช่น การศึกษาของ Fellaji และคณะ (2014) หรือไม่นำตัวบ่งชี้ด้านพฤติกรรมบางอย่าง เช่นพฤติกรรมการบริโภค เข้ามาเป็นส่วนหนึ่งในการสร้างโมเดลเช่น การสร้างโมเดลสำหรับพยากรณ์การเกิดโรคหัวใจและหลอดเลือด (Cardiovascular) โดย Atoui และคณะ (2006) หรือ Twardy และคณะ (2005) ซึ่งอาจไม่สามารถระบุสาเหตุที่แท้จริงของการเกิดโรคต่างๆ ที่มาจากปัจจัยความเสี่ยงทางด้านพฤติกรรมได้

นอกจากนี้ยังพบว่าข้อมูลจากการสำรวจสุขภาพของประชาชนไทยโดยการตรวจร่างกายไม่เคยถูกนำมาใช้เพื่อสร้างโมเดลแสดงความสัมพันธ์ระหว่างปัจจัยเสี่ยงที่เกี่ยวข้องกับโรคไม่ติดต่อเรื้อรังทั้ง 2 โรคดังกล่าวด้วยวิธีเครือข่ายแบบเบย์ โดยสรุปแล้วผลจากงานวิจัยในครั้งนี้จะทำให้ได้โมเดลที่

แสดงความสัมพันธ์ของสาเหตุหรือปัจจัยเสี่ยงของทั้ง 2 โรคที่ครอบคลุมปัจจัยเสี่ยงทางชีวภาพ (Biomarker) ตัวบ่งชี้ด้านพฤติกรรม (Behavioral maker) และตัวบ่งชี้ด้านลักษณะประชากร (Characteristic marker) จากฐานข้อมูลระดับประชากรของประเทศไทย (การสำรวจสุขภาพประชาชนไทยโดยการตรวจร่างกาย ครั้งที่ 4)

วิธีเครือข่ายแบบเบย์ (Bayesian network) มีการจดลิขสิทธิ์ถึงวิธีการสำหรับโมเดลการตัดสินในทางการแพทย์แบบอัตโนมัติ (Automated medical decision) โดยสร้างจากความรู้ทางการแพทย์ เพื่อใช้ในการวินิจฉัย หรือประเมินสถานการณ์ของคนไข้โดย Sadeghi et al. ในปี 2004 นอกจากนี้ยังมีการใช้โมเดลแบบเบย์ไปพัฒนาเป็น Web-based modeling system ที่มีชื่อว่า miniTUBA (Xiang, Minter, Bi, Woolf, & He, 2007) ที่ใช้ชุดของข้อมูลมาทำการวิเคราะห์ทางการแพทย์ หรือ DIAMED (Cléret, Le Duff, Fresnel, & Le Beux, 2001) ที่เป็นระบบการใช้งานผ่านเว็บในการช่วยวินิจฉัยโรค

Ture et al. (Ture, Kurt, Turhankurum, & Ozdamar, 2005) มีงานวิจัยที่พยายามเปรียบเทียบเทคนิคเพื่อทำนายการเกิดโรคความดันโลหิตสูงจากวิธี Learning จากข้อมูลด้วยวิธีต่างๆ เช่น เทคนิคที่เกี่ยวกับต้นไม้การตัดสินใจ (Decision tree) ได้แก่ Chi-squared automatic interaction detector, Classification and regression tree, Quick unbiased efficient statistical tree เทคนิคทางสถิติได้แก่Logistic regression, Flexible discriminant analysis, Multivariate additive regression splinesเทคนิคโครงข่ายประสาทเทียม (Neural networks) เช่น Multi-Layer perception และ Radical basis function additive regression splines และสุดท้ายคือวิธี Hierarchical cluster analysis และจากการผลการเปรียบเทียบเทคนิคดังกล่าวพบว่าเทคนิคด้วยวิธีโครงข่ายประสาทเทียมให้ผลการพยากรณ์การเกิดโรคความดันโลหิตสูงได้ดีที่สุดจากข้อมูลเพื่อใช้ทดสอบผลการเปรียบเทียบเทคนิคต่างๆ มีผลสอดคล้องกับการพยากรณ์การเกิดโรคหัวใจและหลอดเลือด (Coronary artery disease) ที่เสนอผลการวิจัยโดย Kurt et al. (2008) ถึงแม้ว่าวิธีโครงข่ายประสาทเทียม (Neural networks) จะให้ผลการพยากรณ์การเกิดโรคที่ดีว่าวิธีอื่นๆ แต่วิธีนี้ก็มีข้อด้อยตรงที่วิธีนี้เป็น Black box ซึ่งยากที่จะอธิบาย ในข้อด้อยดังกล่าวสามารถแก้ไขได้โยการใช้วิธีเครือข่ายแบบเบย์ (Bayesian network) โครงสร้างของโมเดลที่นำเสนอด้วยรูปกราฟนั้นง่ายที่จะเข้าใจ (Suvisaari et al., 2011) และยังพบว่าจากการศึกษาของ Ture et al. (2005) และ Kurt et al. (2008) ยังไม่ได้นำวิธี เครือข่ายแบบเบย์ (Bayesian network) มาใช้ในการเปรียบเทียบ

สิ่งที่น่าสนใจก็คือจากการศึกษาของ Atoui et al. (2006) พบว่าโมเดลเครือข่ายแบบเบย์สามารถพยากรณ์การเกิดโรคหัวใจและหลอดเลือดจากฐานข้อมูลได้ดีกว่า Neural network และ Logistic regression แต่การศึกษาดังกล่าวไม่สามารถยืนยันได้ว่าโมเดลเครือข่ายแบบเบย์จะสามารถพยากรณ์การเกิดโรคเบาหวานและความดันโลหิตสูงได้ดีกว่าวิธีอื่นโดยเฉพาะการศึกษาในกลุ่มประชาชนไทย

จากการทบทวนวรรณกรรมยังพบว่าวิธีเครือข่ายแบบเบย์เคยนำมาใช้เพื่อวินิจฉัยหรือพยากรณ์การเกิดโรคต่างๆ จากฐานข้อมูลผู้ป่วยที่เข้ารับการรักษาในโรงพยาบาล เช่น การพยากรณ์ความเสี่ยงการเป็นโรคมะเร็งเต้านม (Burnside et al., 2006) การพยากรณ์การเกิดโรคโลหิตจาง (Sebastiani et al., 2007) หรือการประเมินความเสี่ยงต่อสุขภาพจากมลพิษทางอากาศ (Liu et al., 2011) แต่พบว่าการศึกษาการใช้โมเดลเครือข่ายแบบเบย์มาทำนายการเกิดโรคเบาหวานและความดันโลหิตสูงโดยเฉพาะนั้นยังพบไม่แพร่หลาย จากการทบทวนวรรณกรรมพบว่ามีการศึกษาเพื่อพยากรณ์การเกิดเบาหวานชนิดที่ 2 ด้วยโมเดลแบบเบย์ โดย Guo, Bai and Hu (2012) พบว่าโมเดลสามารถพยากรณ์ได้ถูกต้องคิดเป็น 72.3% ส่วนงานวิจัยอื่นๆ ที่พบ เช่น Atoui et al. (2006) หรือ Fellaji, Azmani and Akharif (2014) มีการศึกษาโดยกำหนดให้โรคเบาหวานเป็นหนึ่งในสาเหตุการเกิดโรคอื่นๆ และมิได้มุ่งเน้นการหาสาเหตุของการเกิดโรคเบาหวานโดยตรง

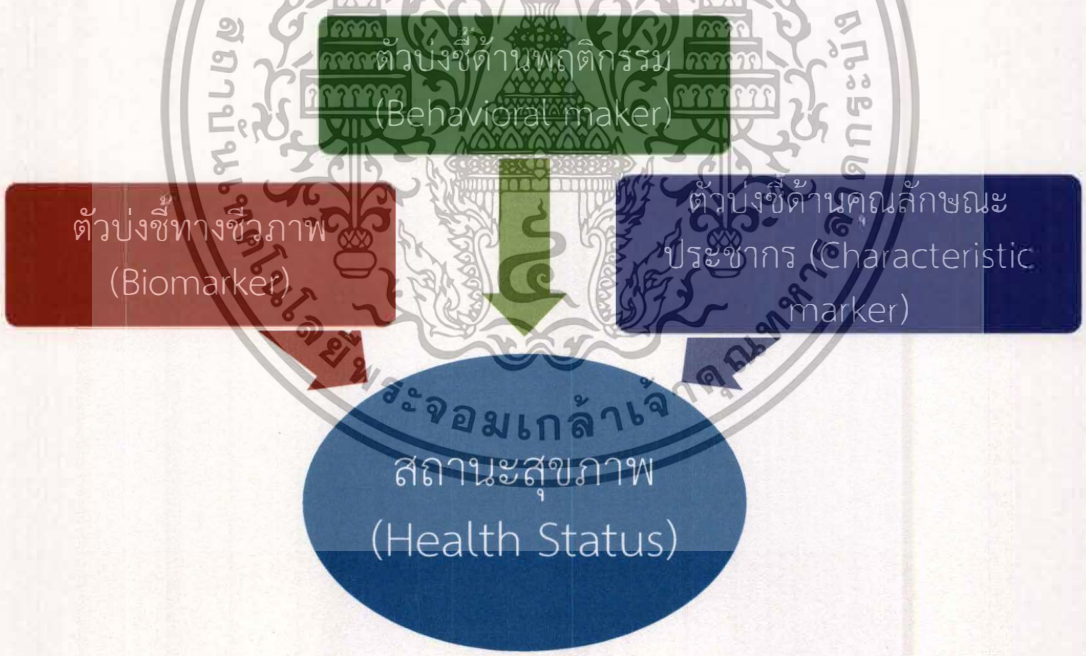
นอกจากนี้ยังพบว่าข้อมูลจากการสำรวจสุขภาพของประชาชนไทยโดยการตรวจร่างกายไม่เคยถูกนำมาใช้เพื่อสร้างโมเดลแสดงความสัมพันธ์ระหว่างปัจจัยเสี่ยงที่เกี่ยวข้องกับโรคไม่ติดต่อเรื้อรังทั้ง 2 โรคดังกล่าวด้วยวิธีเครือข่ายแบบเบย์ โดยสรุปแล้วผลจากงานวิจัยในครั้งนี้จะทำให้ได้โมเดลที่แสดงความสัมพันธ์ของสาเหตุหรือปัจจัยเสี่ยงของทั้ง 2 โรคที่ครอบคลุมปัจจัยเสี่ยงทางชีวภาพ (Biomarker) ตัวบ่งชี้ด้านพฤติกรรม (Behavioral maker) และตัวบ่งชี้ด้านลักษณะประชากร (Characteristic marker) จากฐานข้อมูลระดับประชากรของประเทศไทย (การสำรวจสุขภาพประชาชนไทยโดยการตรวจร่างกาย ครั้งที่ 4)

บทที่ 3 วิธีการและแผนการดำเนินงานวิจัย

ในบทนี้จะนำเสนอแนวทางการดำเนินการวิจัยเพื่อสร้างตัวแบบเครือข่ายแบบเบย์ด้วยวิธีการเรียนรู้ด้วยเครื่อง พร้อมทั้งแนวทางการเปรียบเทียบโมเดลในรูปแบบที่แตกต่างกัน

3.1 กรอบแนวความคิด

จากการทบทวนวรรณกรรม สามารถกำหนดปัจจัยหลักที่ส่งผลต่อสถานะทางสุขภาพของการเกิดโรคไม่ติดต่อเรื้อรัง จากสาเหตุหลัก 3 ตัวบ่งชี้ได้แก่ ตัวบ่งชี้ด้านคุณลักษณะประชากร ตัวบ่งชี้ด้านชีวภาพ และตัวบ่งชี้ด้านพฤติกรรม จึงนำมาใช้เป็นตัวกำหนดกรอบแนวความคิดการวิจัยเพื่อนำมาเป็นกรอบในการกำหนดตัวแบบเครือข่ายแบบเบย์ที่เหมาะสมต่อไป ซึ่งแสดงได้ดัง รูป 3-1



รูป 3-1 กรอบการศึกษาแสดงความสัมพันธ์ของปัจจัยต่างๆ ที่ส่งผลต่อสถานะสุขภาพ

3.2 วิธีดำเนินการวิจัย

การดำเนินงานวิจัยเน้นการนำข้อมูลทุติยภูมิที่มีการเก็บรวบรวมไว้จากการสำรวจสุขภาพประชาชนไทยโดยการตรวจร่างกายครั้งที่ 4 พร้อมด้วยข้อมูลปฐมภูมิจากการสัมภาษณ์แพทย์

ผู้เชี่ยวชาญในสาขาโรคเบาหวาน เพื่อให้แน่ใจว่าโมเดลที่สร้างขึ้นมีความถูกต้องและเหมาะสมตามหลักฐานทางการแพทย์โดยมีขั้นตอนในการสร้างโมเดลดังตาราง 3-1

ตาราง 3-1 สรุปขั้นตอนการสร้างโมเดลเครือข่ายแบบเบย์จากข้อมูลการสำรวจสุขภาพประชาชนไทยโดยการตรวจร่างกาย

ขั้นตอนการสร้างโมเดล
1. ศึกษาวิธีการสร้างเครือข่ายแบบเบย์ด้วยวิธี Structure learning และทบทวนวรรณกรรมที่เกี่ยวข้อง
2. ติดต่อและทำข้อตกลงเพื่อขอใช้ข้อมูลการสำรวจสุขภาพของประชาชนไทยโดยการตรวจร่างกาย
3. สร้างเครือข่ายแบบเบย์ด้วยวิธี Structure learning และวิธีการอื่นๆ โดยแบ่งข้อมูลในอัตรา 70:30 เพื่อกำหนดเป็น Training dataset และ Testing dataset
4. ตรวจสอบความถูกต้องของโครงสร้างโมเดลกับแพทย์ผู้เชี่ยวชาญเฉพาะโรค (โมเดลสามารถปรับเปลี่ยนแก้ไขโครงสร้างตามคำแนะนำของแพทย์ผู้เชี่ยวชาญ 2 ท่าน
5. กำหนดค่าความน่าจะเป็นของแต่ละปัจจัยตามโครงสร้างโมเดลที่สร้างไว้ด้วยวิธี Parameter learning จากฐานข้อมูลที่ได้จากการสำรวจฯ
6. ใช้โมเดลที่มีข้อมูลความน่าจะเป็นเรียบร้อยแล้วมาทำการวิเคราะห์เพื่อพยากรณ์ความชุกของการเกิดโรค
7. ตรวจสอบความถูกต้องสมบูรณ์จากผลการวิเคราะห์ที่ได้จากโมเดลกับแพทย์ผู้เชี่ยวชาญสัมภาษณ์แพทย์ผู้เชี่ยวชาญ
8. ปรับปรุงแก้ไขโมเดลตามคำแนะนำของผู้เชี่ยวชาญ
9. เปรียบเทียบผลที่ได้ด้วยวิธีการเรียนรู้ด้วยเครื่อง (Learning) และวิธีความเห็นของผู้เชี่ยวชาญ (Expert) ตามหลักเกณฑ์ที่กำหนด
10. ทำรายงานฉบับสมบูรณ์ก่อนนำเสนอรายงานผลการวิเคราะห์ไปยังหน่วยงานที่เกี่ยวข้อง

อนึ่งขั้นตอนการสร้างโมเดลได้ได้รับความอนุเคราะห์จาก ศาสตราจารย์นายแพทย์ วิชัย เอกพลากร หัวหน้าภาควิชาเวชศาสตร์ชุมชน คณะแพทยศาสตร์ โรงพยาบาลรามาธิบดี ที่ให้ความเห็นและคำแนะนำในการสร้างโมเดลเครือข่ายแบบเบย์ และอนุเคราะห์ข้อมูลจากฐานข้อมูลการสำรวจสุขภาพของประชาชนไทย จากสำนักงานการสำรวจสภาวะสุขภาพของประชาชนไทย (สสท.)

ข้อมูลที่รวบรวมมาได้ในแต่ละขั้นตอนนำมาวิเคราะห์ที่ภาควิชาสถิติ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ด้วยโปรแกรมพื้นฐานทางสถิติที่มีอยู่ และโปรแกรมเฉพาะทางเพื่อสร้างโมเดลเครือข่ายแบบเบย์ซึ่งเน้นโปรแกรมที่ไม่เสียค่าใช้จ่ายเพื่อลดปัญหาการเข้าถึงโมเดลโดยผู้ใช้เพื่อลดเงื่อนไขการเข้าโปรแกรมและโมเดล แต่เนื่องจากงานวิจัยนี้ต้องอาศัยการสร้างโมเดลโดยการเรียนรู้จากเครื่องสำหรับโมเดลเครือข่ายแบบเบย์ ที่ต้องอาศัยซอฟแวร์เฉพาะมีจำนวนไม่แพร่หลาย และส่วนใหญ่เป็นซอฟแวร์เพื่อการค้าอันเป็นอุปสรรคสำคัญในการทำวิจัยนี้ ด้วย

ข้อจำกัดด้านงบประมาณผู้วิจัยได้เลือกซอฟแวร์ที่ใช้ในงานวิจัยนี้ซึ่งประกอบด้วย WEKA version 3.7, GeNIE version 2.0 และ SPSS version 19 นำมาวิเคราะห์และแสดงการศึกษาในครั้งนี้

3.3 การกำหนดโครงสร้างโมเดลที่ใช้ในการเปรียบเทียบ

สำหรับการสร้างโมเดลเครือข่ายแบบเบย์นั้นมีทั้งการใช้ผู้เชี่ยวชาญ การเรียนรู้จากเครื่องแบบอัตโนมัติ หรือการรวมเอาทั้ง 2 วิธีนั้นไว้ด้วยกัน (Julia Flores, Nicholson, Brunskill, Korb, & Mascaro, 2011) เพื่อให้เข้าใจถึงความสามารถของโมเดลเครือข่ายแบบเบย์ต่อความสามารถในการพยากรณ์การเกิดโรคเบาหวาน ที่มีวิธีการสร้างโมเดลที่แตกต่างกัน จึงทำการศึกษาโมเดลที่สร้างจากทั้ง 2 กลุ่มวิธี คือโมเดลที่สร้างจากกลุ่มผู้เชี่ยวชาญ และที่สร้างมาจากการเรียนรู้ของเครื่อง

I.	*ความเห็นจากผู้เชี่ยวชาญ*	1. BNE (BN_Expert)
II.	*การเรียนรู้จากเครื่อง*	2. NLT (BN_Learning with Top-down)
		3. BNLB (BN_Learning with Bottom-up)

นอกจากนี้เพื่อให้เข้าใจถึงอิทธิผลของโครงสร้างโมเดลที่มีลักษณะการแสดงผลความสัมพันธ์ในรูปแบบ Hierarchical และ Non Hierarchical ที่ส่งผลต่อความสามารถของโมเดล จึงได้มีการสร้างโมเดลที่เป็น Non-hierarchical มาทำการเปรียบเทียบ อีก 2 โมเดลคือ

III.	*โมเดลที่เป็น Non-hierarchical*	4. *แบบง่าย* (S: BN_Simple)
		5. แบบง่ายลดตัวแปร (SR: BN_SimReduce)

สุดท้ายเป็นการเปรียบเทียบโมเดลแผนภาพต้นไม้การตัดสินใจเทียบกับโมเดลเครือข่ายแบบเบย์ จึงเพิ่มการสร้างโมเดลแผนภาพต้นไม้เพื่อการตัดสินใจซึ่งเป็นการเรียนรู้ด้วยเครื่อง เป็นตัวแบบที่ 6

IV.	*โมเดลแผนภาพต้นไม้เพื่อการตัดสินใจ*	6. DTL (DT_Learning)

3.4 การวิเคราะห์จากตัวแบบเครือข่ายแบบเบย์

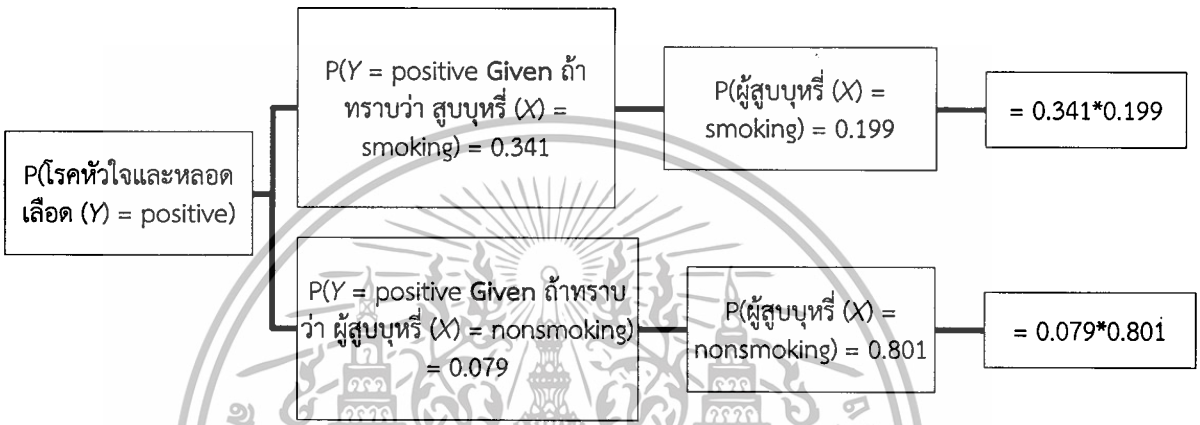
ในการประยุกต์ใช้โมเดลเครือข่ายแบบเบย์กับการพยากรณ์การเกิดโรคเบาหวานมีขั้นตอนและรายละเอียดของวิธีการ จะได้นำเสนอในหัวข้อต่อไปนี้

3.4.1 การหาความน่าจะเป็นของการเกิดโอกาสที่ไม่พึงประสงค์

การสร้างโมเดลเป็นการกำหนดความสัมพันธ์ของเหตุการณ์ไม่พึงประสงค์ที่อาจก่อให้เกิดผลต่อตัวแปรผลลัพธ์ที่สนใจทั้งที่เป็นสาเหตุทางตรง (Direct effect) และทางอ้อม (Indirect effect) ซึ่งเป็นโมเดลโครงสร้างเชิงเหตุและผล (Cause-Effect diagram) หากตัวแปรที่เป็นตัวแปรสาเหตุย่อย (Child variable) ในขั้นตอนการกำหนดค่าความน่าจะเป็นถูกกำหนดโดยตาราง CPT ซึ่งไม่ใช่การกำหนดค่าความน่าจะเป็นของแต่ละสถานะ (State) ของตัวแปรโดยตรง ดังนั้นหากต้องการหาค่าความน่าจะเป็นของการเกิดเหตุการณ์ไม่พึงประสงค์ต่างๆ ที่อยู่ในรูปตัวแปรสาเหตุย่อย (Child

variable) หรือตัวแปรผลกระทบ (Effect variable) ต้องใช้ทฤษฎีเครือข่ายแบบเบย์เพื่อคำนวณค่าดังกล่าว

ตัวอย่าง จากตัวอย่างข้างต้นการกำหนดค่าความน่าจะเป็นของเหตุการณ์หรือตัวแปร 'การสูบบุหรี่' ที่ปรากฏใน (Probability Table: PT) แต่ค่าความน่าจะเป็นของการเกิด 'โรคหัวใจและหลอดเลือด' ไม่สามารถกำหนดลงในในโมเดลได้โดยตรงในตาราง CPT ซึ่งต้องอาศัยหลักการคำนวณจาก 0.341*0.199 + 0.079*0.801 = 0.131138, ซึ่งแสดงได้ดังรูป 3-2



รูป 3-2 การคำนวณโอกาสการเกิด 'โรคหัวใจและหลอดเลือด' (Y)

รายละเอียดแสดงการคำนวณสามารถแสดงได้ดังนี้

$$P(Y = \text{positive}) = \sum_{x} P(Y = \text{positive}|X)P(X)$$
$$= P(Y = \text{positive}|X = \text{smoking}) * P(X = \text{smoking})$$
$$+ P(Y = \text{positive}|X = \text{nonsmoking}) * P(X = \text{nonsmoking})$$
$$= 0.341 * 0.199 + 0.079 * 0.801 = 0.067859 + 0.063279 = 0.131138$$

ความน่าจะเป็นที่จะไม่เป็นโรคหัวใจและหลอดเลือด คือ:

$$P(Y = \text{negative}) = 1 - P(Y = \text{positive}) = 1 - 0.131138 = 0.868862$$

โดยทั่วไปการกำหนดความน่าจะเป็นมีส่วนสำคัญของตัวแปร (ทั้งสาเหตุหรือผลกระทบตัวแปร) ในรูปความน่าจะเป็นของการเกิดขึ้นสำหรับแต่ละสถานะ (State) ของตัวแปรโดยเฉพาะ ซึ่งเรียกว่า "a marginal prior" โดยในรายงานนี้จะเรียกว่า "ความน่าจะเป็นในปัจจุบัน" ของเหตุการณ์

ไม่พึงประสงค์ที่สนใจ ความน่าจะเป็นดังกล่าวใช้เพื่ออธิบายความเป็นไปได้ของเหตุการณ์ไม่พึงประสงค์ที่จะเกิดขึ้นของแต่ละตัวแปรในโมเดล

อนึ่งเฉพาะตัวแปรปัจจัยที่เป็นต้นเหตุของปัญหา (Root cause) ซึ่งเป็นตัวแปรต้นกำเนิดของการเกิดโรคในโมเดล จะมีความน่าจะเป็นเท่ากับความน่าจะเป็นที่ป้อนข้อมูล (Input) ในตาราง PT แต่สำหรับตัวแปรตัวแปรที่ไม่ใช่ตัวแปรต้นเหตุ หรือเรียกว่า 'Child variable' หรือผลลัพธ์ที่มีผลต่อความน่าจะเป็นของแต่ละตัวแปรที่จะเกิดขึ้น (Marginal probability) ที่ไม่ใช่ข้อมูลนำเข้า (Input) โดยตรงและจะได้จากการคำนวณโดยสมการ (2-1)

ตัวอย่างแสดงความสัมพันธ์ของตัวแปร 2 ตัวแปรเมื่อตัวแปร X เป็นสาเหตุของตัวแปร Y ดังรูป 3-3

รูป 3-3 ตัวอย่างโมเดลเครือข่ายแบบเบย์ซึ่งแสดงความสัมพันธ์ของตัวแปร 1 คู่

การคำนวณโอกาสของการเกิด Y สามารถคำนวณได้จากสมการ (3-1)

$$P(Y)=\sum_{x} P(Y|X)P(X) \qquad \textbf{(3-1)}$$

3.4.2 การวิเคราะห์ด้วยโมเดลเครือข่ายแบบเบย์

การวิเคราะห์โมเดลที่นำมาใช้ในการเปรียบเทียบความสามารถในการพยากรณ์จำแนกกลุ่มผู้เป็นโรคและไม่เป็นโรคเบาหวานนั้น ผู้วิจัยได้ทำการเปรียบเทียบโมเดลที่สร้างขึ้นด้วยวิธีการดังนี้

1. ผลการวิเคราะห์ด้วยเส้นโค้ง ROC และ AUC
2. การวิเคราะด้วยค่า TP Rate, FP Rate, TN Rate, FN Rate, Precision, Recall, F1, Accuracy และ G-mean

บทที่ 4 ผลการวิจัย

ในบทนี้จะนำเสนอผลจากการวิจัยที่ได้จากการสร้างโมเดลเครือข่ายแบบเบย์ที่มีโครงสร้างที่แตกต่างกัน นำไปสู่การเปรียบเทียบความสามารถในการพยากรณ์และการจำแนกกลุ่มผู้ป่วยโรคเบาหวาน และสามารถนำผลที่ได้จากโมเดลนำสู่แนวทางการสนับสนุนการวิเคราะห์ความเสี่ยงของการเกิดโรคไม่ติดต่อที่สำคัญ

ก่อนที่จะกล่าวถึงตัวแปรและนิยาม (หัวข้อ 4.1) ลักษณะทั่วไปของผู้เข้าร่วมโครงการการสำรวจสุขภาพของประชาชนไทยโดยการตรวจร่างกายครั้งที่ 4 (หัวข้อ 4.2) และเพื่อให้เข้าใจผลจากการเปรียบเทียบโมเดลรูปแบบต่างๆ ที่เป็นวัตถุประสงค์หลักของงานวิจัยนี้ จากนั้นจะกล่าวถึงผลจากโครงสร้างโมเดลทั้ง 6 โมเดล (ในหัวข้อ 4.3) และผลการเปรียบเทียบการพยากรณ์ด้วยโมเดลทั้ง 6 (ในหัวข้อ 4.4)

4.1 ตัวแปรและนิยาม

ตัวแปรที่ใช้ในการสร้างเครือข่ายแบบเบย์เพื่อศึกษาการเกิดโรคเบาหวานกำหนดขึ้นโดยการสัมภาษณ์ แพทย์ที่เข้าร่วมในโครงการวิจัยในฐานะผู้เชี่ยวชาญ (Experts) และนายแพทย์ผู้ดำเนินการโครงการการเก็บข้อมูลการสำรวจสุขภาพประชาชนไทยโดยการตรวจร่างกาย ครั้งที่ 4 กำหนดตัวแปรสาเหตุการเกิดโรคเบาหวานที่เกี่ยวข้องทั้งสิ้น 7 ตัวแปร และตัวแปรที่ระบุการเกิดโรคเบาหวานที่ได้มาจากผลการตรวจทางห้องปฏิบัติการ ดังนั้นจึ้งนำตัวแปรทั้ง 8 ตัวแปร มาทำการศึกษา

เริ่มต้นจากการกำหนดตัวแปรเป้าหมาย (Focal Variable) ซึ่งก็คือ การเกิดโรคเบาหวาน (Diabetes) โดยประกอบด้วยตัวแปรที่เป็นสาเหตุของการเกิดโรคเบาหวานทั้งทางตรงและทางอ้อมทั้งสิ้น 7 ตัวแปร โดยกำหนดนิยามของแต่ละตัวแปรจากรายงานการสำรวจสุขภาพของประชาชนไทยโดยการตรวจร่างกาย ครั้งที่ 4 แสดงดังตาราง 4-1

ตาราง 4-1 ตัวแปร นิยาม และกลุ่มระดับตัวแปร (State)

ลำดับ	ตัวแปรและระดับ	นิยาม
1	Age	กลุ่มอายุ
	15-34	อายุระหว่าง 15-34 ปี
	35-59	อายุระหว่าง 35-59 ปี
2	Area of residence	เขตที่พักอาศัย
	Urban	ในเขตเทศบาล
	Rural	นอกเขตเทศบาล
3	Diabetes	ภาวะการเป็นโรคเบาหวาน
	Positive	การตรวจเลือดหลังอดอาหารนาน 12 ชั่วโมง (Fasting Plasma Glucose, FPG) ≥ 126 mg/dl หรือ เคยได้รับการบอกว่าเป็นโรคเบาหวาน
	Negative	การตรวจเลือดหลังอดอาหารนาน 12 ชั่วโมง (Fasting Plasma Glucose, FPG) < 126 mg/dl
4	Family History DM	ประวัติการเป็นโรคเบาหวานของครอบครัว (สายตรง)
	Yes	มีญาติสายตรงได้แก่ พ่อ แม่ หรือพี่น้องมีประวัติเป็นโรคเบาหวาน
	No	ไม่มีญาติสายตรงได้แก่ พ่อ แม่ หรือพี่น้องมีประวัติเป็นโรคเบาหวาน
5	Fruit and Vegetable consumption หรือ Veg	การบริโภคผักและผลไม้
	Less	ผักผลไม้ไม่เพียงพอ < 5 servings/day
	Normal to high	> 5 servings/day
6	Obesity	ภาวะโรคอ้วน
	Normal	ทั้ง BMI และรอบเอวต่ำกว่าเกณฑ์
	Over BMI	BMI ≥ 25 (รอบเอวไม่เกินเกณฑ์)
7	Socio-economic	ความมั่นคงทางเศรษฐกิจและสังคม
	Quantile1	ตัวแปรดัชนี Wealth index score คำนวณ อยู่ใน Quantile 1
	Quantile2	ตัวแปรดัชนี Wealth index score คำนวณ อยู่ใน Quantile 2
	Quantile3	ตัวแปรดัชนี Wealth index score คำนวณ อยู่ใน Quantile 3
	Quantile4	ตัวแปรดัชนี Wealth index score คำนวณ อยู่ใน Quantile 4
	Quantile5	ตัวแปรดัชนี Wealth index score คำนวณ อยู่ใน Quantile 5
8	Physical activity	การเคลื่อนไหวร่างกายที่มีการใช้พลังงานในร่างกายโดยคำนึงถึงระดับความหนักเบาของกิจกรรมทางกาย
	Low	ระดับของการมีกิจกรรมทางกายต่ำกว่าเกณฑ์ระดับปานกลางและระดับมาก
	Medium	a. มีกิจกรรมทางกายไม่มากถึงระดับมาก และ a. มีกิจกรรมอย่างหนัก ≥ 3 วัน/สัปดาห์ และเวลา ≥ 20 นาทีต่อวัน b. มีกิจกรรมปานกลางหรือเดิน รวม ≥ 5 วัน อย่างน้อยวันละ 30 นาทีต่อวัน และ total MET-นาทีต่อสัปดาห์ ≥ 1,500 หรือ c. มีกิจกรรมทางกายอย่างหนักและปานกลางหรือเดิน รวม ≥ 5 วัน/สัปดาห์ และ total MET-นาทีต่อสัปดาห์ ≥ 600
	High	มีกิจกรรมทางกายอย่างหนัก ≥ 3 วัน/สัปดาห์ และ total MET-นาทีต่อสัปดาห์ ≥ 1,500 หรือ มีกิจกรรมทางกายอย่างหนักหรือปานกลางรวม ≥ 7 วัน/สัปดาห์ และ total MET-นาทีต่อสัปดาห์ ≥ 3,000

4.2 ลักษณะของหน่วยตัวอย่างผู้เข้าร่วมโครงการ การสำรวจสุขภาพของประชาชนไทยโดยการตรวจร่างกายครั้งที่ 4

ข้อมูลทั้งหมดที่สำรวจได้จากผู้ที่เข้าร่วมโครงการ การสำรวจสุขภาพของประชาชนไทยโดยการตรวจร่างกายครั้งที่ 4 ที่มีอายุ 15-59 ปี มีจำนวนทั้งสิ้น 11,240 คน ถูกนำมาใช้ในการสร้างโมเดลในงานวิจัยนี้ แต่เนื่องจากมีข้อมูลสูญหายบางตัวแปรทำให้มีผลรวมของความถี่ในการจำแนกตามตัวแปรสนใจแตกต่างกันไป

ตาราง 4-2 แสดงจำนวนและร้อยละของหน่วยตัวอย่างที่เข้าร่วมโครงการฯ โดยพบว่าหน่วยตัวอย่างส่วนใหญ่มีอายุ 35-59 ปี คิดเป็นร้อยละ 66.54 และอาศัยอยู่ในเมืองมากกว่าชนบทเล็กน้อย คิดเป็นร้อยละ 54.50 นอกจากนี้ยังพบว่าประมาณ 1 ใน 3 ของหน่วยตัวอย่างมีญาติสายตรงที่เป็นโรคเบาหวาน (26.17%) หน่วยตัวอย่างส่วนใหญ่บริโภคผักและผลไม้ต่อวันน้อยกว่าเกณฑ์ที่ควรบริโภค (79.87%) และยังพบว่าในกลุ่มหน่วยตัวอย่างมีร้อยละ 37.37 มีน้ำหนักเกินหรือเป็นโรคอ้วน และมีเพียงร้อยละ 19.93 ที่มีกิจกรรมทางกายค่อนข้างต่ำ

ยิ่งไปกว่านี้ยังพบว่า มีผู้ที่ป่วยที่มีภาวะโรคเบาหวานซึ่งเป็นผลการตรวจทางห้องปฏิบัติการพบว่ามีระดับน้ำตาลเกินกว่าเกณฑ์ที่กำหนด จำนวนทั้งสิ้น 598 คน คิดเป็นร้อยละ 5.32 ของหน่วยตัวอย่างทั้งหมด

ตาราง 4-2 ลักษณะของหน่วยตัวอย่าง

ตัวแปร	จำนวน	ร้อยละ
กลุ่มอายุ (Age)		
15-34	3,761	33.46
35-59	7,479	66.54
รวม	**11,240**	**100.00**
เขตที่พักอาศัย (Area of residence)		
Rural	5,114	45.50
Urban	6,126	54.50
รวม	**11,240**	**100.00**
ประวัติการเป็นโรคเบาหวานของครอบครัว (สายตรง) (Family History DM)		
No	8,299	73.83
Yes	2,941	26.17
รวม	**11,240**	**100.00**
การบริโภคผักและผลไม้ (Fruit and Vegetable consumption)		
Less	8,891	79.87
Normal to high	2,241	20.13
รวม	**11,132**	**100.00**

ตาราง 4-2 (ต่อ)

ตัวแปร	จำนวน	ร้อยละ
ภาวะโรคอ้วน (Obesity)		
Normal	7,022	62.63
Over	4,190	37.37
รวม	**11,212**	**100.00**
ความมั่นคงทางเศรษฐกิจและสังคม (Socio-economic)		
Quantile1	1,635	14.55
Quantile2	1,839	16.36
Quantile3	2,497	22.22
Quantile4	2,214	19.70
Quantile5	3,055	27.18
รวม	**11,240**	**100.00**
การเคลื่อนไหวร่างกาย (Physical activity)		
Low	2,218	19.93
Medium	2,720	24.44
High	6,192	55.63
รวม	**11,130**	**100.00**
การเป็นโรคเบาหวาน (Diabetes)		
Positive	598	5.32
Negative	10,642	94.68
รวม	**11,240**	**100.00**

4.3 การแบ่งข้อมูลที่ใช้ในการศึกษา

จากข้อมูลของของหน่วยตัวอย่างทั้งหมดจำ 11,240 นำมาแบ่งเป็น 2 กลุ่ม ในอัตรา 70:30 อย่างสุ่ม เรียกกลุ่มแรกว่า Training dataset มีทั้งสิ้นจำนวน 8,003 คน และกลุ่มที่ 2 เรียกว่า Testing dataset จำนวน 3,237 คน โดยจำแนกจำนวนผู้เป็นโรคเบาหวานและผู้ไม่เป็นโรคเบาหวาน แสดงดังตาราง 4-3 ในรูปจำนวนและร้อยละของผู้ป่วยโรคเบาหวาน พบว่าทั้ง 2 กลุ่มข้อมูลมีสัดส่วนผู้มีภาวะการเกิดโรคเบาหวานใกล้เคียงกัน คิดเป็นร้อยละ 5.24 และร้อยละ 5.53 จากข้อมูล Training และ Testing dataset ซึ่งแสดงให้เห็นว่าการแบ่งข้อมูลออกเป็น 2 กลุ่มเกิดขึ้นอย่างสุ่ม

ตาราง 4-3 จำนวนและร้อยละของตัวอย่างจำแนกตามภาวะโรคเบาหวานที่พบในแต่ละกลุ่มข้อมูล

สถานะการเป็นโรคเบาหวาน	Training Dataset		Testing Dataset	
	ความถี่	%	ความถี่	%
Positive	419	5.24	179	5.53
Negative	7,584	94.76	3,058	94.47
รวม	**8,003**	**100.00**	**3,237**	**100.00**

4.4 โครงสร้างโมเดลที่ใช้ในการศึกษา

การศึกษาเปรียบเทียบโมเดลในการพยากรณ์การเกิดโรคเบาหวานสำหรับงานวิจัยนี้ ประกอบด้วยโมเดลทั้ง 6 โมเดล ตามเทคนิค ลักษณะโครงสร้าง และวิธีการที่แตกต่างกัน ดังตาราง 4-4

ตาราง 4-4 คุณลักษณะเฉพาะของโมเดลที่นำมาศึกษา

เทคนิคที่ใช้	ลักษณะโครงสร้าง	วิธีการ	โมเดล
BN	Hierarchical	Expert	1. BNE (BN_Expert)
BN	Hierarchical	Learning	2. BNLB (BN_Learning with Bottom-up)
BN	Hierarchical	Learning	3. BNLT (BN_Learning with Top-down)
BN	Non-Hierarchical	Simple	4. S (BN_Simple)
BN	Non-Hierarchical	Simple	5. SR (BN_SimReduce)
DT	Hierarchical	Learning	6. DTL (DT_Learning

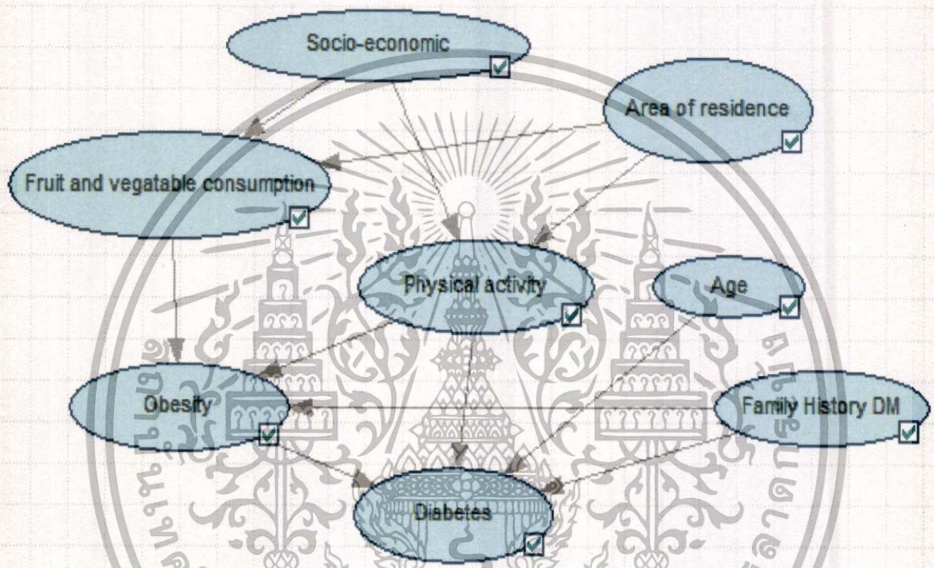
จากการกำหนดโมเดลที่ใช้ในการศึกษาทั้ง 6 โมเดล พบว่า โมเดลที่ 1-5 เป็นโมเดลที่สร้างขึ้นด้วยเทคนิคโมเดลเครือข่ายแบบเบย์ ส่วนโมเดลที่ 6 คือโมเดลที่สร้างขึ้นด้วนเทคนิคต้นไม้ตัดสินใจ

โมเดลที่ 1, 2, 3 และ 6 มีโครงสร้างแบบลำดับขั้น (Hierarchical Structure) ส่วนโมเดลที่ 4 และ 5 เป็นโมเดลที่ไม่มีโครงสร้างลำดับขั้นของโมเดล (Non-hierarchical Structure)

โมเดลที่ 2, 3 และ 6 สร้างขึ้นด้วยวิธีการเรียนรู้ด้วยเครื่อง ในขณะที่โมเดลที่ 1 สร้างขึ้นจากความคิดเห็นของผู้เชี่ยวชาญ ส่วนโมเดลที่ 4 และ 5 สร้างขึ้นโดยวิธีการอย่างง่าย

4.4.1 โครงสร้างโมเดลเครือข่ายแบบเบย์โดยผู้เชี่ยวชาญ (BNE)

การสร้างโมเดลเครือข่ายแบบเบย์แบบแรกได้จากความคิดเห็นแพทย์ที่เข้าร่วมในโครงการวิจัยในฐานนะผู้เชี่ยวชาญ (Experts) และนายแพทย์ผู้ดำเนินการโครงการการเก็บข้อมูลการสำรวจสุขภาพประชาชนไทยโดยการตรวจร่างกาย ครั้งที่ 4 ร่วมกำหนดโครงสร้างในเชิงเหตุและผลเพื่อแสดงสาเหตุของการเกิดโรคเบาหวานทั้งในรูปสาเหตุทางตรงและทางอ้อม โมเดลผ่านการตรวจสอบและปรับปรุงหลายครั้งจนกระทั่งได้รูปแบบโครงสร้างของเครือข่ายแบบเบย์ที่เหมาะสมและมีข้อมูลจากการสำรวจฯ สนับสนุนการคำนวณค่าความน่าจะเป็นจากข้อมูล Training dataset จนกระทั่งได้โครงสร้างโมเดล ซึ่งแสดงได้ดังต่อไปนี้



รูป 4-1 โครงสร้างโมเดลเครือข่ายแบบเบย์โดยผู้เชี่ยวชาญสำหรับโรคเบาหวาน (BNE)

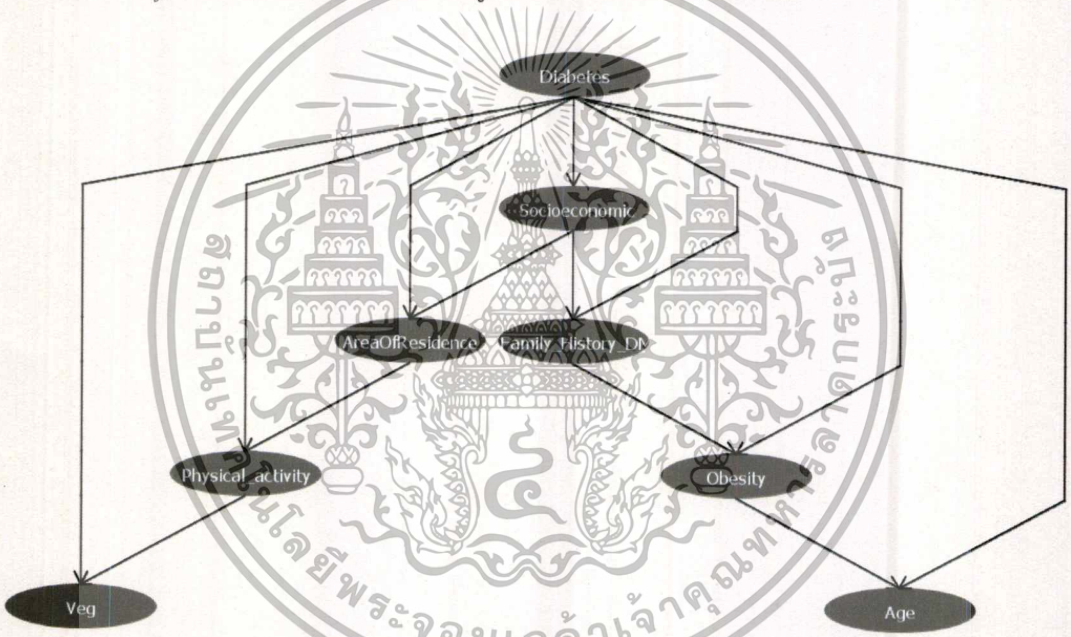
โมเดลเครือข่ายแบบเบย์โดยผู้เชี่ยวชาญ (BNE) สำหรับโรคเบาหวาน เป็นโมเดลที่สร้างขึ้นที่มีลักษณะโครงสร้างแบบเป็นลำดับขั้น แสดงความสัมพันธ์เชิงเหตุและผลของการเกิดโรคเบาหวาน ในรูปสาเหตุทางตรงและทางอ้อม โดยการทำ workshop กับผู้เชี่ยวชาญที่ได้ดำเนินการไว้แล้วในงานวิจัยก่อนหน้านี้ (Leerojanaprapa et al., 2017) ความสัมพันธ์ระหว่างตัวแปรถูกกำหนดขึ้นโดยเชื่อมตัวแปรที่อาจส่งผลต่อการเกิดโรคเบาหวานไปยังตัวแปรภาวะการเกิดโรคเบาหวานจึงถูกกำหนดให้เป็นตัวแปรสุดท้ายที่มีตัวแปรอื่นเชื่อมลูกศรไปสู่ตัวแปรดังกล่าว เนื่องจากสาเหตุการเกิดโรคนั้นมีได้หลากหลาย จากนั้นมีการกำหนดระดับของแต่ละตัวแปรตามที่กำหนดไว้ในหัวข้อ 4.1 โดยความคิดเห็นของผู้เชี่ยวชาญประกอบกับข้อมูลที่มีมีอยู่ในชุดข้อมูลที่ได้จากการสำรวจสุขภาพของประชาชนไทยโดยการตรวจร่างกายที่เป็นข้อมูลหลักสำหรับนำมากำหนดค่าความน่าจะเป็นของแต่ละตัวแปรจากชุดข้อมูล Training dataset

4.4.2 โมเดลเครือข่ายแบบเบย์ด้วยวิธีเรียนรู้จากเครื่องแบบบนลงล่าง (BNLT)

โมเดลที่ 2 คือโมเดลเครือข่ายแบบเบย์ที่ได้จากการเรียนรู้ด้วยเครื่องโดยใช้ข้อมูล Training dataset มีลักษณะแบบลำดับขั้น ผู้วิจัยใช้โปรแกรม Weka 3.6 ในการสร้างโมเดลนี้โดยใช้ Search Algorithm TAN ซึ่ง คือ Tree Augmented Naïve Bayes ซึ่งถูกสร้างขึ้นโดยการคำนวณค่า Maximum weight spanning tree โดยใช้ Crow and Lui algorithm ที่อ้างไว้ในโดย Bouckaert (2008) หลังจากที่โมเดลได้ถูกสร้างขึ้นแล้ว ขั้นต่อไปเป็นการเรียนรู้ของเครื่องเพื่อกำหนดค่าความน่าจะเป็นของแต่ละตัวแปร ในที่นี้เลือกวิธี Simple Estimator ที่กำหนดค่าความน่าจะเป็นแบบมีเงื่อนไขจากสูตรต่อไปนี้

$$P(x_i = k \mid pa(x_i) = j) = \frac{N_{ijk} + N'_{ijk}}{N_{ij} + N'_{ij}}$$

เมื่อ N'_{ijk} คือ alpha parameter ที่ถูกกำหนดใน default เป็น 0.5

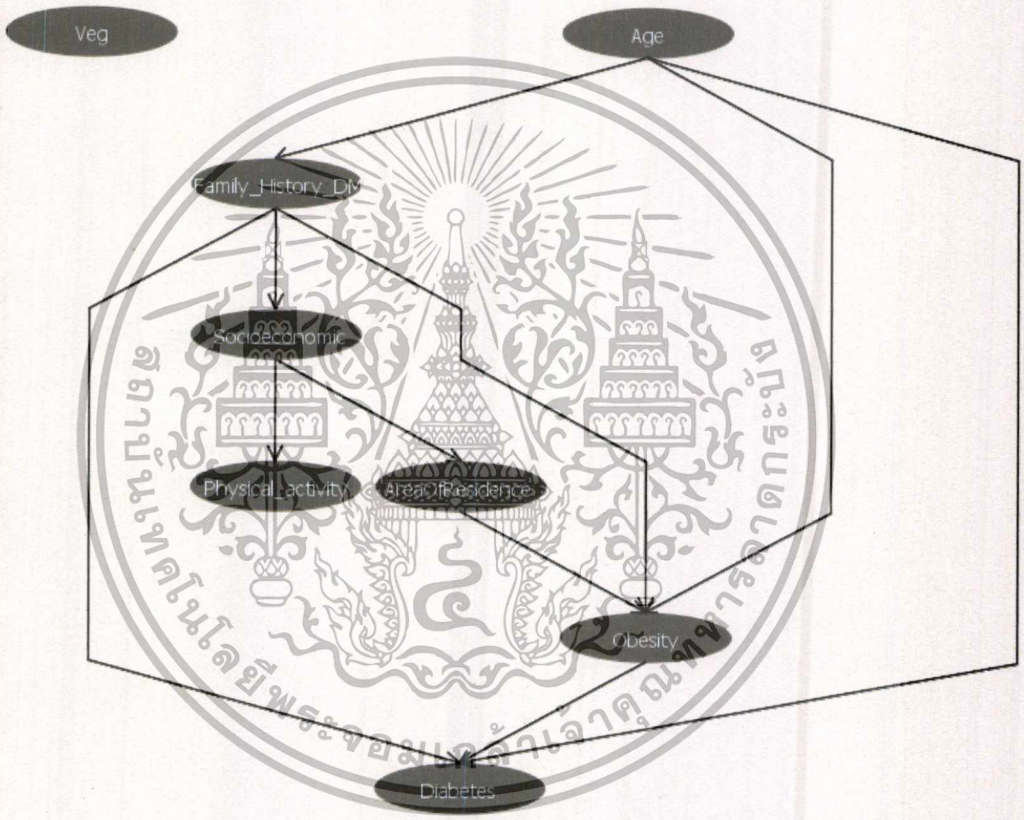


รูป 4-2 โครงสร้างโมเดลเครือข่ายแบบเบย์โดยการเรียนรู้จากเครื่องแบบบนลงล่าง (BNLT)

จากรูป 4-2 จะเห็นได้ว่า ตัวแปร Diabetes เป็นตัวแปรที่อยู่สูงสุดที่ไม่มีตัวแปรใดมีลูกศรชี้เข้าตัวแปรดังกล่าว ซึ่งมักเรียกว่า Root node และยังพบว่ามีทิศทางของลูกศรเชื่อมไปยังตัวแปรอื่นๆ เป็นลำดับขั้น และมีตัวแปร Veg (การบริโภคผักและผลไม้, Fruit and Vegetable consumption) และ Age (อายุ) เป็นตัวแปรที่อยู่ปลายสุด ซึ่งมักจะเรียกว่า Leaf node

4.4.3 โมเดลเครือข่ายแบบเบย์ด้วยวิธีเรียนรู้จากเครื่องแบบล่างขึ้นบน (BNLB)

โมเดลที่ 3 นี้ คือโมเดลเครือข่ายแบบเบย์ที่ได้จากการเรียนรู้ด้วยเครื่องโดยใช้ข้อมูล Training dataset มีลักษณะแบบลำดับขั้น เช่นเดียวกับโมเดลที่ 2 ผู้วิจัยใช้โปรแกรม Weka 3.7 ในการสร้างโมเดลนี้เช่นกัน แต่ใช้ Search Algorithm ชื่อ Generic Search ในการกำหนดโครงสร้างของโมเดลที่อ้างไว้ในโดย Bouckaert (2008) หลังจากที่โมเดลได้ถูกสร้างขึ้นแล้ว ขั้นต่อไปเป็นการเรียนรู้ของเครื่องเพื่อกำหนดค่าความน่าจะเป็นของแต่ละตัวแปร ในที่นี้เลือกวิธี Simple Estimator เช่นเดียวกับโมเดลที่ 2



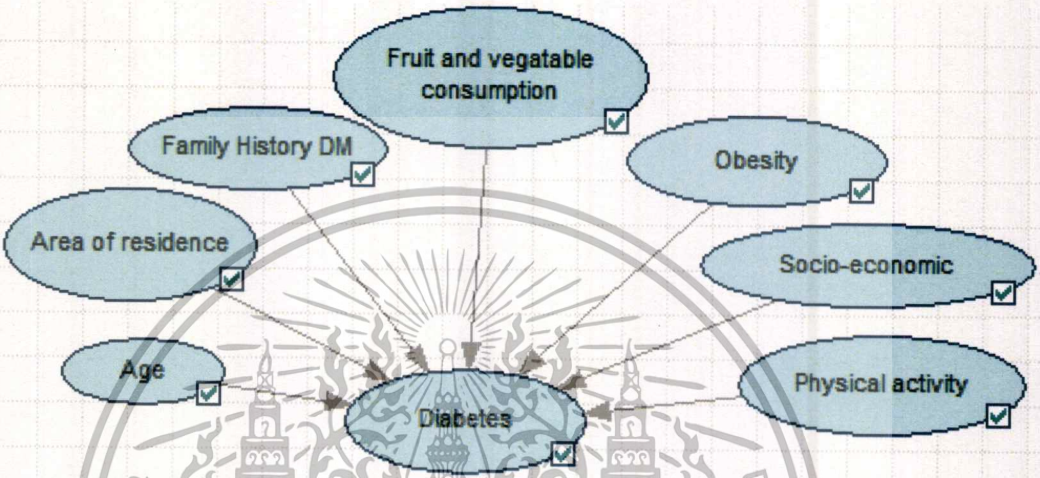
รูป 4-3 โครงสร้างโมเดลเครือข่ายแบบเบย์โดยการเรียนรู้จากเครื่องแบบล่างขึ้นบน (BNLB)

จากรูป 4-3 จะเห็นได้ว่า ตัวแปร Diabetes เป็นตัวแปรที่อยู่ล่างสุดที่ไม่มีลูกศรชี้ออกจากตัวแปรนี้แล้ว ซึ่งมักจะเรียกว่า Leaf node และยังพบว่ามีทิศทางของลูกศรเชื่อมไปยังตัวแปรอื่นๆ เป็นลำดับขั้น และมีตัวแปร Age (อายุ) เป็นตัวแปรที่อยู่ปลายสุด ซึ่งมักเรียกว่า Root node

นอกจาหนี้ยังพบว่าตัวแปร Physical activity (การเคลื่อนไหวร่างกาย) เป็นอีกตัวแปรที่เป็น Leaf node แต่ไม่ส่งผลต่อตัวแปร Diabetes จากโมเดลยังชี้ให้เห็นว่าตัวแปร Veg (การบริโภคผักและผลไม้, Fruit and Vegetable consumption) เป็นตัวแปรที่ไม่มีความสัมพันธ์ต่อการเกิดโรคเบาหวานทั้งทางตรงและทางอ้อม

4.4.4 โมเดลเครือข่ายแบบเบย์อย่างง่าย (S)

โมเดลแบบง่ายสร้างขึ้นโดยไม่คำนึงถึงโครงสร้างลำดับขั้นของโมเดล (Non-hierarchical Structure) โดยกำหนดให้ทุกตัวแปรสาเหตุสัมพันธ์โดยตรงกับการเกิดโรคเบาหวาน แสดงได้ดังรูป 4-4 สำหรับการกำหนดค่าความน่าจะเป็นในแต่ละตัวแปรใช้ข้อมูลจาก Training dataset เป็นตัวกำหนด



รูป 4-4 โมเดลเครือข่ายแบบเบย์อย่างง่าย (S)

4.4.5 โมเดลเครือข่ายแบบเบย์ลดตัวแปรอย่างง่าย (SR)

โมเดลนี้สร้างขึ้นโดยไม่คำนึงถึงโครงสร้างลำดับขั้นของโมเดล พัฒนาจากโมเดลเครือข่ายแบบเบย์อย่างง่ายที่นำเสนอในหัวข้อ 4.4.4 แต่เพื่อให้มั่นใจว่าตัวแปรที่มีลูกศรเชื่อมสู่ตัวแปร Diabetes มีความสัมพันธ์กับการเกิดโรคเบาหวาน จึงทำการทดสอบความสัมพันธ์ระหว่างตัวแปรและเลือกเฉพาะตัวแปรที่มีความสัมพันธ์กับการเกิดโรคเบาหวานอย่างมีนัยสำคัญมาสร้างโมเดลในรูปแบบง่าย หากตัวแปรใดไม่สัมพันธ์กับตัวแปร Diabetes ก็จะตัดตัวแปรดังกล่าวออกไปจากโมเดลเพื่อเป็นการพัฒนาโมเดลให้ดียิ่งขึ้น ในเบื้องต้นจึงให้การทดสอบสมมุติฐานเพื่อทดสอบความเป็นอิสระระหว่างตัวแปร Diabetes กับตัวแปรที่เหลือทีละคู่ โดยใช้สถิติ Pearson Chi-Square test of independence ผลการทดสอบแสดงดังตารางที่ 4-5

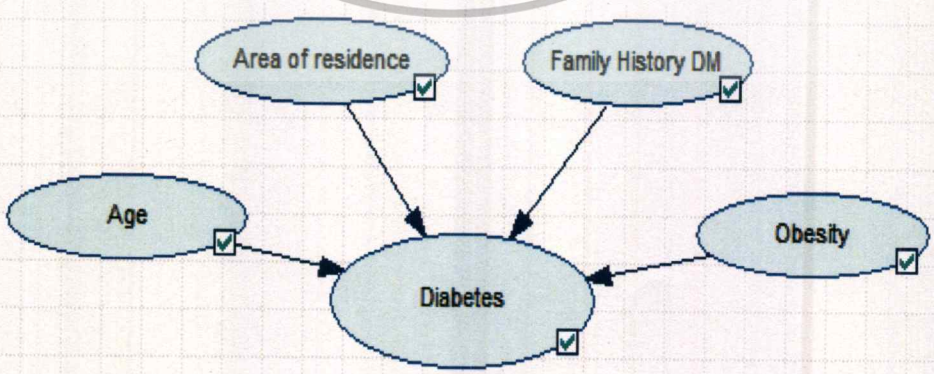
ตาราง 4-5 Pearson Chi-Square และ p-value สำหรับการทดสอบความเป็นอิสระระหว่างตัวแปร Diabetes และตัวแปรอื่นๆ

ตัวแปร	Pearson Chi-Square	p-value
Age	22.3382	0.0000*
Area of residence	199.2318	0.0000*
Family History DM	266.1512	0.0000*
Fruit and vegetable Consumption	1.5101	0.2191
Obesity	188.1879	0.0000*
Socio-economic	4.8520	0.3028
Physical activity	0.9525	0.3291

* p-value < 0.05

ผลจากการทดสอบพบว่า เมื่อกำหนดระดับนัยสำคัญที่ 0.05 มีตัวแปรที่มีความสัมพันธ์กับการเกิดโรคเบาหวาน 4 ตัวแปร ซึ่งประกอบด้วย Age, Area of residence, Family History DM และ Obesity เพราะเนื่องจากมีค่า p-value < 0.05 (ค่า p-value = Exact Sig. (2-sided)) ซึ่งแสดงว่า ตัวแปรทั้งคู่ขึ้นอยู่ต่อกัน สำหรับตัวแปร Veg: Fruit and vegetable Consumption, Socio-economic, และ Physical activity มีค่า p-value ≥ 0.05 แสดงให้เห็นว่าตัวแปรทั้ง 3 ไม่มีความสัมพันธ์ต่อการเกิดโรคเบาหวานอย่างมีนัยสำคัญทางสถิติ

จากผลการทดสอบที่ได้ข้างต้น จึงนำเฉพาะตัวแปรที่มีความสัมพันธ์อย่างมีนัยสำคัญมาสร้างโมเดลเครือข่ายแบบเบย์อย่างง่ายโดยไม่คำนึงโครงสร้างลำดับชั้น แสดงได้ดังรูป 4-5

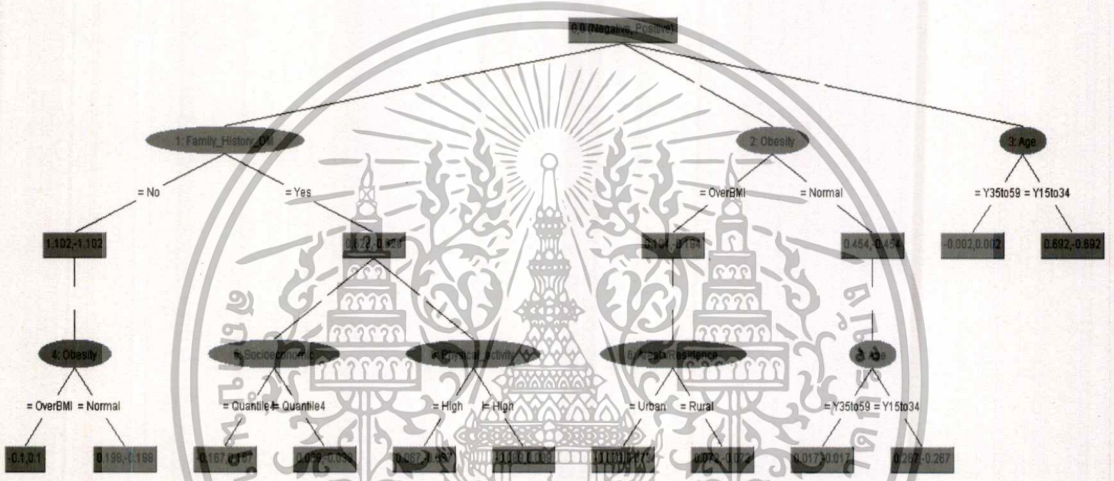


รูป 4-5 โมเดลเครือข่ายแบบเบย์ลดตัวแปรแบบง่ายแบบ (SR)

4.4.6 โมเดลต้นไม้ตัดสินใจด้วยวิธีเรียนรู้จากเครื่อง (DTL)

โมเดลสุดท้ายเป็นโมเดลที่ไม่ใช่รูปแบบของโมเดลเครือข่ายแบบเบย์ แต่มักนิยมใช้ในการพยากรณ์การจำแนกกลุ่มเช่นเดียวกับโมเดลเครือข่ายแบบเบย์ ที่เรียกว่าโมเดลต้นไม้ตัดสินใจ จะเห็นได้ว่าโมเดลต้นไม้ในการตัดสินใจมีลักษณะโครงสร้างแบบลำดับชั้น และสร้างขึ้นโดยวิธีการเรียนรู้จากเครื่องจากชุดข้อมูล Training dataset ชุดเดียวกับการสร้างโมเดลเครือข่ายแบบเบย์ทั้งโมเดลที่ 2 และ 3

ในการสร้างโมเดลต้นไม้ตัดสินใจจากการเรียนรู้จากเครื่องนี้ ผู้วิจัยใช้โปรแกรม Weka 3.6 ในการสร้างโมเดลนี้เช่นกัน แต่ใช้ LAD Tree ซึ่งเป็นการสร้างโมเดลต้นไม้การตัดสินใจที่ใช้ LogitBoost strategy ในการกำหนดโครงสร้างของโมเดล โมเดลต้นไม้ตัดสินใจแสดได้ดังรูป 4-6



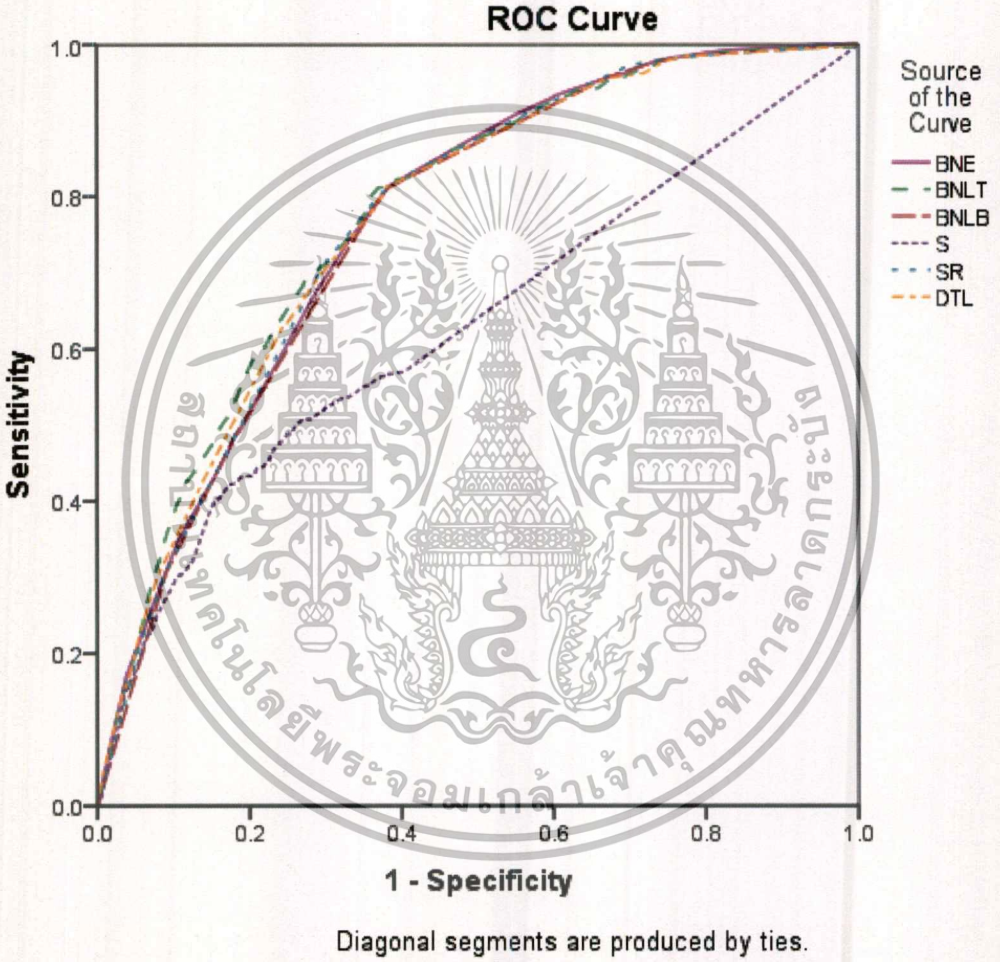
รูป 4-6 โมเดลต้นไม้ตัดสินใจ (DTL)

4.5 ผลเปรียบเทียบพยากรณ์ด้วยโมเดลต่างๆ

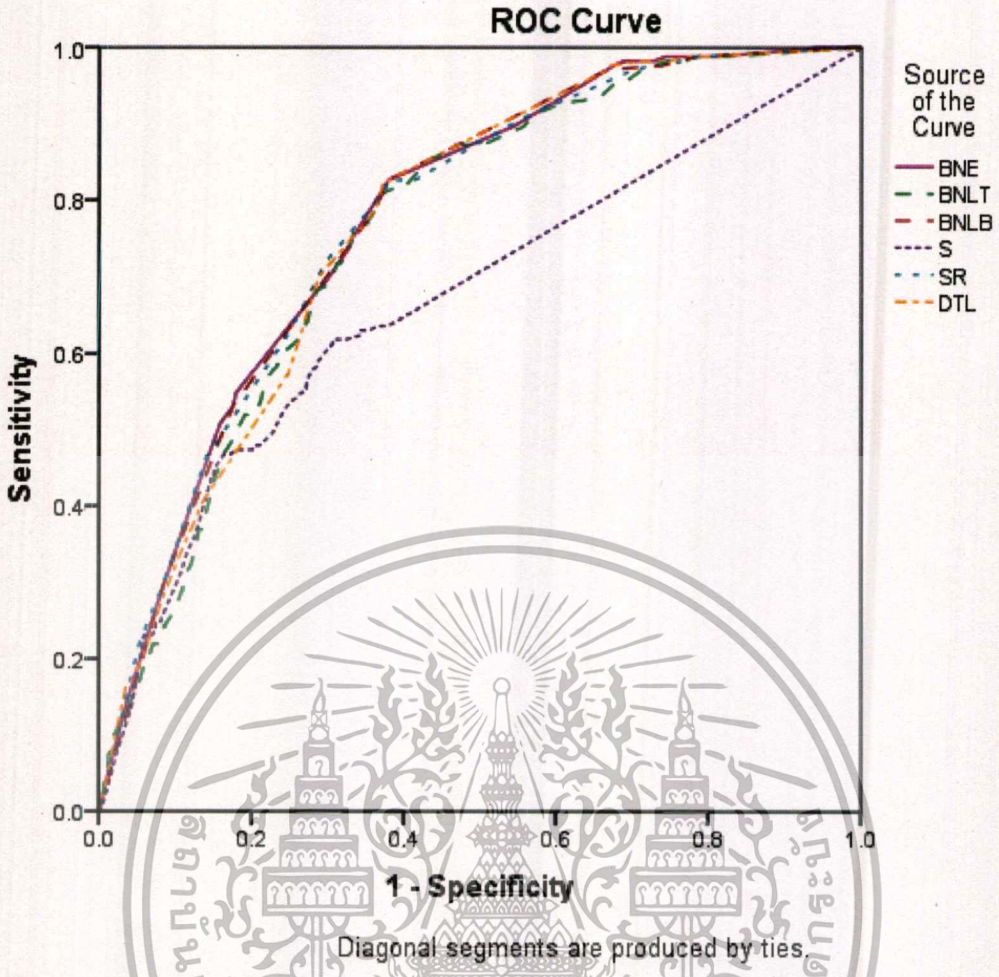
การเปรียบเทียบความสามารถของโมเดลทั้ง 6 โมเดลนั้น โดยสามารถทำการเปรียบเทียบได้โดยพิจารณาผลการวิเคราะห์จากเส้นโค้ง ROC และพื้นที่ใต้เส้นโค้ง AUC และผลจากการวิเคราะห์ด้วยคอนฟูชั่น แมททริก ซึ่งจะได้แสดงผลการวิเคราะห์ในแต่ละหัวข้อย่อยต่อไปนี้

4.5.1 ผลการวิเคราะห์ด้วยเส้นโค้ง ROC และ AUC

สำหรับการประเมินความสามารถในการจำแนกกลุ่มผู้ป่วยโรคเบาหวานที่ได้จากการพยากรณ์ความน่าจะเป็นของการเกิดโรคเบาหวานด้วยโมเดลแต่ละโมเดลสามารถทำการเปรียบเทียบได้ด้วยเส้นโค้ง Receiver Operating Characteristic (ROC) curves และ the area under curve (AUC) ที่ได้จากการประเมินจาก Training dataset และ Testing dataset แสดงดังรูป 4-7 – 4-8 และตาราง 4-6



รูป 4-7 เส้นโค้ง ROC สำหรับโมเดลการจำแนกกลุ่มทั้ง 6 โมเดล ที่สร้างจาก Training dataset



รูป 4-8 เส้นโค้ง ROC สำหรับโมเดลการจำแนกกลุ่มทั้ง 6 โมเดล ที่สร้างจาก Testing dataset

ตาราง 4-6 ค่า AUC จาก Training และ Testing dataset

Model	Training Dataset	ลำดับ	Testing Dataset	ลำดับ
BNE	0.7670	2	0.7765	1
BNLT	0.7764	1	0.7612	5
BNLB	0.7607	5	0.7749	2
S	0.6281	6	0.6685	6
SR	0.7667	4	0.7741	3
DTL	0.7694	3	0.7689	4

ผลจากเส้นโค้ง ROC ที่ได้จาก Training dataset แสดงดังรูป 4-7 และเส้นโค้ง ROC ที่ได้จาก Testing dataset แสดงดังรูป 4-8 มีลักษณะใกล้เคียงกัน โดยพบว่า โมเดล BNE, BNLT, BNLB SR และ DTL มีรูปแบบเส้นโค้ง ROC ที่มีลักษณะใกล้เคียงกันและเข้าใกล้มุมบนซ้ายมือ ยกเว้นโมเดล S ที่มีเส้นโค้งเข้าใกล้เส้นทแยงมุม โดยพบว่า เส้นโค้ง ROC ของโมเดล BNLT ดีกว่าโมเดลอื่นๆ เมื่อพิจารณาจากข้อมูล Training dataset ในขณะที่โมเดล BNE ดีกว่าโมเดลอื่นๆ เมื่อพิจารณาจากข้อมูล Testing dataset

ผลจากเส้นโค้ง ROC สอดคล้องกับค่า AUC ซึ่งแสดงดังตาราง 4-6 พบว่าสำหรับข้อมูล Training dataset โมเดล BNLT มีค่า AUC สูงสุด รองลงมาคือโมเดล BNE และอันดับ 3 คือ DTL ตามด้วย SR, BNLB และ S ตามลำดับ สำหรับข้อมูล Testing dataset พบว่าโมเดล BNE มีค่า AUC สูงสุด รองลงมาคือโมเดล BNLB และอันดับ 3 คือ SR ตามด้วย DTL, BNLT และ S ตามลำดับ จะเห็นได้ว่าค่า AUC ที่ได้จากแต่ละโมเดล จาก Training dataset และ Testing dataset มีค่าและลำดับที่ไม่คงที่ แต่เมื่อพิจารณาในภาพรวมแล้ว BNE เป็นโมเดลที่ให้ค่า AUC สูงกว่าโมเดลอื่น เนื่องจากอยู่ในลำดับที่ 2 ในกลุ่ม Training dataset และลำดับที่ 1 จากกลุ่ม Testing dataset

4.5.2 ผลการวิเคราะห์ด้วยคอนฟูชั่น แมททริก

คอนฟูชั่น แมททริกแสดงการจำแนกการเกิดโรคเบาหวานของแต่ละโมเดล โดยนำ Testing dataset มาทำการทดสอบโดยใช้โมเดลที่สร้างขึ้นจาก Training dataset ที่ได้กล่าวไว้ในหัวข้อ 4.4 โดยกำหนดระดับเกณฑ์การแบ่งกลุ่มที่แตกต่างกัน ซึ่งในที่นี้กำหนดในระดับความน่าจะเป็นระหว่าง 0.1-0.5 เพื่อใช้ในการจำแนกกลุ่มผู้ที่เป็นโรคเบาหวานและไม่เป็นโรคเบาหวานจากการพยากรณ์ในรูปความน่าจะเป็นของการเกิดโรคเบาหวานที่ได้จากการวิเคราะห์ด้วยโมเดลที่ใช้ในการเปรียบเทียบทั้ง 6 โมเดล แสดงผลการจำแนกกลุ่มดังตารางที่ 4-7

ตาราง 4-7 คอนฟูชั่น แมททริก (Confusion matrix) และร้อยละการจำแนกผิดกลุ่ม จำแนกตามชนิดโมเดล และระดับเกณฑ์ในการแบ่งกลุ่ม (Threshold)

โมเดล	Actual	ระดับเกณฑ์ในการแบ่งกลุ่ม (Threshold)									
		0.5		0.4		0.3		0.2		0.1	
		Predict		Predict		Predict		Predict		Predict	
		P	N	P	N	P	N	P	N	P	N
BNE	P	0	173	0	173	0	173	11	162	92	81
	N	0	2,890	0	2,890	0	2,890	44	2,864	507	2,383
	ER	5.65%		5.65%		5.65%		6.73%		19.20%	
BNLT	P	0	173	0	173	0	173	18	155	90	83
	N	0	2,890	0	2,890	0	2,890	65	2,825	553	2,337
	ER	5.65%		5.65%		5.65%		7.18%		20.76%	
BNLB	P	0	173	0	173	0	173	0	173	92	81
	N	0	2,890	0	2,890	0	2,890	0	2,890	507	2,383
	ER	5.65%		5.65%		5.65%		5.65%		19.20%	
S	P	23	150	23	150	36	137	54	119	86	87
	N	119	2,771	120	2,770	162	2,728	310	2,580	660	2,230
	ER	8.78%		8.81%		9.76%		14.01%		24.39%	
SR	P	0	173	0	173	0	173	0	173	121	52
	N	0	2,890	0	2,890	0	2,890	0	2,890	828	2,062
	ER	5.65%		5.65%		5.65%		5.65%		28.73%	
DTL	P	0	173	0	173	0	173	12	161	70	103
	N	0	2,890	0	2,890	0	2,890	45	2,845	391	2,499
	ER	5.65%		5.65%		5.65%		6.73%		16.13%	

P: Positive; N: Negative; ER (Error Rate)

ผลจากการวิเคราะห์ด้วยตารางคอนฟูชั่น แมททริก ทำให้ทราบว่า เมื่อลดระดับเกณฑ์ในการแบ่งกลุ่มลดลงทำให้ค่าความคลาดเคลื่อนในการพยากรณ์เพิ่มขึ้น แต่ทำให้ผลบวกจริง (TP) และผลบวกลวง (FP) เพิ่มมากขึ้น ในขณะที่ผลลบจริง (TN) และผลลบลวง (FN) ลดลง

ผลการจำแนกด้วยคอนฟูชั่น แมททริกเมื่อกำหนดเกณฑ์ที่ 0.5 ซึ่งเป็นเกณฑ์ที่สูงที่สุด พบว่าข้อมูลจากโมเดลเกือบทุกโมเดลยกเว้นโมเดล SR ไม่มีการพยากรณ์ว่าเป็นโรคเบาหวานเลยเนื่องจากไม่มีข้อมูลของหน่วยตัวอย่างใดที่ให้ค่าความน่าจะเป็นในการพยากรณ์การเกิดโรคเกินค่า 0.5 จึงทำให้เกิดเพียงผลลบลวง (FN) และผลลบจริง (TN) ทำให้เกิดค่าความคลาดเคลื่อนในการพยากรณ์คิดเป็น 5.65%

เมื่อเริ่มเปลี่ยนระดับเกณฑ์การแบ่งกลุ่มเป็น 0.2 พบว่า โมเดล BNE, BNLT, S และ DTL ให้ค่าในตารางคอนฟูชั่น แมททริก เปลี่ยนไปจากระดับเกณฑ์การแบ่งกลุ่มที่สูงกว่าดังกล่าว และพบว่าที่ระดับการแบ่งกลุ่มเป็น 0.2 โมเดล BNE และ DTL ให้ผลการจำแนกที่ดีใกล้เคียงกันที่ค่าความ

คลาดเคลื่อนในการพยากรณ์คิดเป็น 6.73% ตามด้วยโมเดล BNTL ซึ่งมีค่าความคลาดเคลื่อนในการพยากรณ์คิดเป็น 7.18% และ โมเดล S ซึ่งมีค่าความคลาดเคลื่อนในการพยากรณ์คิดเป็น 14.01% ตามลำดับ

ผลการจำแนกด้วยคอนฟูชั่น แมททริกเปลี่ยนแปลงทุกโมเดลเมื่อระดับเกณฑ์การแบ่งกลุ่มเป็น 0.1 พบว่า DTL ซึ่งมีค่าความคลาดเคลื่อนในการพยากรณ์ต่ำที่สุด คิดเป็น 16.13% ตามมาด้วยโมเดล BNE และ BNLB ซึ่งมีค่าความคลาดเคลื่อนในการพยากรณ์เท่ากันคิดเป็น 19.20% และตามด้วย โมเดล BNLT, S และ SR ตามลำดับ

เมื่อพิจารณาค่าที่ใช้ตรวจสอบความสามารถในพยากรณ์ของตัวแบบ โดยใช้เกณฑ์ ทั้ง 9 เกณฑ์ ซึ่งประกอบด้วย TP Rate, FP Rate, TN Rate, FN Rate, Precision, Recall, F1, Accuracy, และ G-mean ของทุกโมเดล ดังตารางที่ 4-8 พบว่า เมื่อลดค่าระดับเกณฑ์ในการแบ่งกลุ่มลง ค่า TP Rate และ FP Rate เพิ่มขึ้นในขณะที่ TN Rate และ FN Rate ลดลง นอกจากนี้ยังพบว่ามีค่า Accuracy มีแนวโน้มลดลงอีกด้วย

เมื่อพิจารณา ค่า Precision พบว่ามีแนวโน้มลดลง ขณะที่ค่า Recall มีแนวโน้มเพิ่มขึ้นเมื่อลดค่าระดับเกณฑ์การแบ่งกลุ่มในเกือบทุกโมเดล (ยกเว้นโมเดล BNLB และ SR ที่ไม่สามารถบอกทิศทางแนวโน้มของค่า Precision ได้ เนื่องจากสามารถคำนวณค่า Precision ได้เฉพาะที่เกณฑ์ระดับในการแบ่งกลุ่มที่ 0.1 เท่านั้น) และด้วยเพราะมีค่า Precision และค่า Recall มีทิศทางสวนทางกัน จึงต้องนำมาพิจารณาจากค่า F1 โดยพบว่าค่า F1 มีแนวโน้มเพิ่มขึ้นเมื่อค่าระดับเกณฑ์การแบ่งกลุ่มลง ยกเว้นโมเดล S ที่ให้ค่า F1 สูงที่สุดเมื่อระดับเกณฑ์การแบ่งกลุ่มเป็น 0.2

เมื่อพิจารณาค่า G-mean ยังพบว่ามีแนวโน้มเพิ่มขึ้นเมื่อมีระดับเกณฑ์ในการแบ่งกลุ่มลดลง ซึ่งมีทิศทางเดียวกับค่า F1

ตาราง 4-8 เกณฑ์ที่สำคัญในการเปรียบเทียบรายโมเดล จำแนกตามระดับเกณฑ์ในการแบ่งกลุ่ม (Threshold levels)

โมเดล	ระดับเกณฑ์ในการแบ่งกลุ่ม (Threshold)					แนวโน้ม
	0.5	0.4	0.3	0.2	0.1	
BNE						
TP Rate	0.0000	0.0000	0.0000	0.0636	0.5318	↑
FP Rate	0.0000	0.0000	0.0000	0.0151	0.1754	↑
TN Rate	1.0000	1.0000	1.0000	0.9849	0.8246	↓
FN Rate	1.0000	1.0000	1.0000	0.9364	0.4682	↓
Precision	-	-	-	0.2000	0.1536	↓
Recall	0.0000	0.0000	0.0000	0.0636	0.5318	↑
F1	-	-	-	0.0965	0.2383	↑
Accuracy	0.9435	0.9435	0.9435	0.9331	0.8080	↓
G-mean	0.0000	0.0000	0.0000	0.2502	0.6622	↑
BNLT						
TP Rate	0.0000	0.0000	0.0000	0.1040	0.5202	↑
FP Rate	0.0000	0.0000	0.0000	0.0225	0.1913	↑
TN Rate	1.0000	1.0000	1.0000	0.9775	0.8087	↓
FN Rate	1.0000	1.0000	1.0000	0.8960	0.4798	↓
Precision	-	-	-	0.2169	0.1400	↓
Recall	0.0000	0.0000	0.0000	0.1040	0.5202	↑
F1	-	-	-	0.1406	0.2206	↑
Accuracy	0.9435	0.9435	0.9435	0.9282	0.7924	↓
G-mean	0.0000	0.0000	0.0000	0.3189	0.6486	↑
BNLB						
TP Rate	0.0000	0.0000	0.0000	0.0000	0.5318	↑
FP Rate	0.0000	0.0000	0.0000	0.0000	0.1754	↑
TN Rate	1.0000	1.0000	1.0000	1.0000	0.8246	↓
FN Rate	1.0000	1.0000	1.0000	1.0000	0.4682	↓
Precision	-	-	-	-	0.1536	⊗
Recall	0.0000	0.0000	0.0000	0.0000	0.5318	↑
F1	-	-	-	-	0.2383	⊗
Accuracy	0.9435	0.9435	0.9435	0.9435	0.8080	↓
G-mean	0.0000	0.0000	0.0000	0.0000	0.6622	↑

ตาราง 4-8 (ต่อ)

โมเดล	ระดับเกณฑ์ในการแบ่งกลุ่ม (Threshold)					
	0.5	0.4	0.3	0.2	0.1	
S						
TP Rate	0.1329	0.1329	0.2081	0.3121	0.4971	↑
FP Rate	0.0412	0.0415	0.0561	0.1073	0.2284	↑
TN Rate	0.9588	0.9585	0.9439	0.8927	0.7716	↓
FN Rate	0.8671	0.8671	0.7919	0.6879	0.5029	↓
Precision	0.1620	0.1608	0.1818	0.1484	0.1153	↓
Recall	0.1329	0.1329	0.2081	0.3121	0.4971	↑
F1	0.1460	0.1456	0.1941	0.2011	0.1872	↓
Accuracy	0.9122	0.9119	0.9024	0.8599	0.7561	↓
G-mean	0.3570	0.3570	0.4432	0.5279	0.6193	↑
SR						
TP Rate	0.0000	0.0000	0.0000	0.0000	0.6994	↑
FP Rate	0.0000	0.0000	0.0000	0.0000	0.2865	↑
TN Rate	1.0000	1.0000	1.0000	1.0000	0.7135	↓
FN Rate	1.0000	1.0000	1.0000	1.0000	0.3006	↓
Precision	-	-	-	-	0.1275	⊗
Recall	0.0000	0.0000	0.0000	0.0000	0.6994	↑
F1	-	-	-	-	0.2157	⊗
Accuracy	0.9435	0.9435	0.9435	0.9435	0.7127	↓
G-mean	0.0000	0.0000	0.0000	0.0000	0.7064	↑
DTL						
TP Rate	0.0000	0.0000	0.0000	0.0694	0.4046	↑
FP Rate	0.0000	0.0000	0.0000	0.0156	0.1353	↑
TN Rate	1.0000	1.0000	1.0000	0.9844	0.8647	↓
FN Rate	1.0000	1.0000	1.0000	0.9306	0.5954	↓
Precision	-	-	-	0.2105	0.1518	↓
Recall	0.0000	0.0000	0.0000	0.0694	0.4046	↑
F1	-	-	-	0.1043	0.2208	↑
Accuracy	0.9435	0.9435	0.9435	0.9327	0.8387	↓
G-mean	0.0000	0.0000	0.0000	0.2613	0.5915	↑

ในการเปรียบเทียบระหว่างโมเดลในแต่ละเกณฑ์การจำแนกกลุ่มแสดงดังตาราง 4-9 และรูป 4-9 – 4-17 ได้ผลดังต่อไปนี้

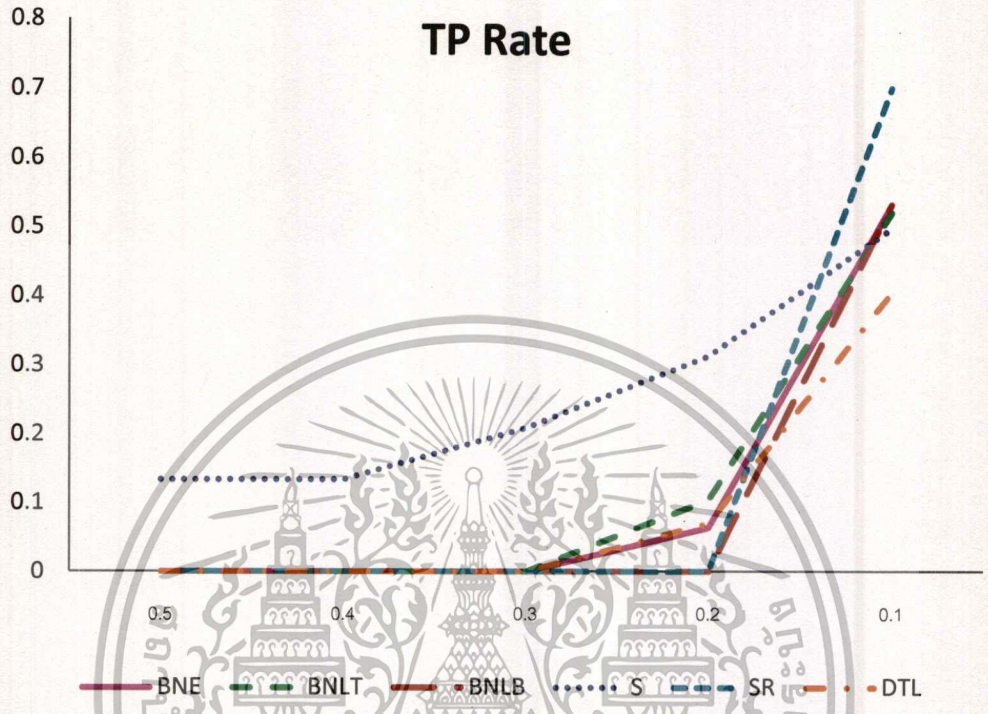
ตาราง 4-9 เกณฑ์ที่สำคัญในการเปรียบเทียบโมเดล จำแนกตามระดับเกณฑ์ในการแบ่งกลุ่ม (Threshold levels) และชนิดของโมเดล

เกณฑ์การตรวจสอบ	ระดับเกณฑ์ในการแบ่งกลุ่ม (Threshold)				
	0.5	0.4	0.3	0.2	0.1
TP Rate					
BNE	0.0000	0.0000	0.0000	0.0636	0.5318
BNLT	0.0000	0.0000	0.0000	0.1040	0.5202
BNLB	0.0000	0.0000	0.0000	0.0000	0.5318
S	0.1329	0.1329	0.2081	0.3121	0.4971
SR	0.0000	0.0000	0.0000	0.0000	0.6994
DTL	0.0000	0.0000	0.0000	0.0694	0.4046
FP Rate					
BNE	0.0000	0.0000	0.0000	0.0151	0.1754
BNLT	0.0000	0.0000	0.0000	0.0225	0.1913
BNLB	0.0000	0.0000	0.0000	0.0000	0.1754
S	0.0412	0.0415	0.0561	0.1073	0.2284
SR	0.0000	0.0000	0.0000	0.0000	0.2865
DTL	0.0000	0.0000	0.0000	0.0156	0.1353
TN Rate					
BNE	1.0000	1.0000	1.0000	0.9849	0.8246
BNLT	1.0000	1.0000	1.0000	0.9775	0.8087
BNLB	1.0000	1.0000	1.0000	1.0000	0.8246
S	0.9588	0.9585	0.9439	0.8927	0.7716
SR	1.0000	1.0000	1.0000	1.0000	0.7135
DTL	1.0000	1.0000	1.0000	0.9844	0.8647
FN Rate					
BNE	1.0000	1.0000	1.0000	0.9364	0.4682
BNLT	1.0000	1.0000	1.0000	0.8960	0.4798
BNLB	1.0000	1.0000	1.0000	1.0000	0.4682
S	0.8671	0.8671	0.7919	0.6879	0.5029
SR	1.0000	1.0000	1.0000	1.0000	0.3006
DTL	1.0000	1.0000	1.0000	0.9306	0.5954

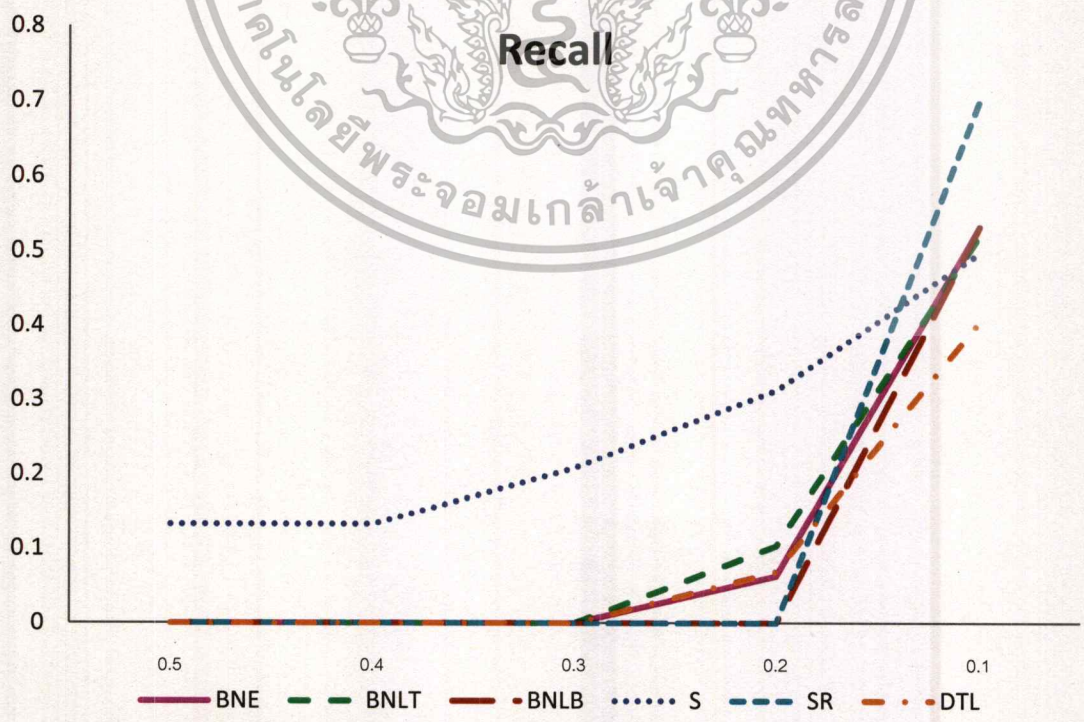
ตาราง 4-9 (ต่อ)

เกณฑ์การตรวจสอบ	ระดับเกณฑ์ในการแบ่งกลุ่ม (Threshold)				
	0.5	0.4	0.3	0.2	0.1
Precision					
BNE	-	-	-	0.2000	0.1536
BNLT	-	-	-	0.2169	0.1400
BNLB	-	-	-	-	0.1536
S	0.1620	0.1608	0.1818	0.1484	0.1153
SR	-	-	-	-	0.1275
DTL	-	-	-	0.2105	0.1518
Recall					
BNE	0.0000	0.0000	0.0000	0.0636	0.5318
BNLT	0.0000	0.0000	0.0000	0.1040	0.5202
BNLB	0.0000	0.0000	0.0000	0.0000	0.5318
S	0.1329	0.1329	0.2081	0.3121	0.4971
SR	0.0000	0.0000	0.0000	0.0000	0.6994
DTL	0.0000	0.0000	0.0000	0.0694	0.4046
F1					
BNE	-	-	-	0.0965	0.2383
BNLT	-	-	-	0.1406	0.2206
BNLB	-	-	-	-	0.2383
S	0.1460	0.1456	0.1941	0.2011	0.1872
SR	-	-	-	-	0.2157
DTL	-	-	-	0.1043	0.2208
Accuracy					
BNE	0.9435	0.9435	0.9435	0.9331	0.8080
BNLT	0.9435	0.9435	0.9435	0.9282	0.7924
BNLB	0.9435	0.9435	0.9435	0.9435	0.8080
S	0.9122	0.9119	0.9024	0.8599	0.7561
SR	0.9435	0.9435	0.9435	0.9435	0.7127
DTL	0.9435	0.9435	0.9435	0.9327	0.8387
G-mean					
BNE	0.0000	0.0000	0.0000	0.2502	0.6622
BNLT	0.0000	0.0000	0.0000	0.3189	0.6486
BNLB	0.0000	0.0000	0.0000	0.0000	0.6622
S	0.3570	0.3570	0.4432	0.5279	0.6193
SR	0.0000	0.0000	0.0000	0.0000	0.7064
DTL	0.0000	0.0000	0.0000	0.2613	0.5915

จากรูป 4-9 และรูป 4-10 ทำให้ทราบว่าค่า TP Rate และ Recall ซึ่งเป็นค่าเดียวกันและมีแนวโน้มเพิ่มขึ้นเมื่อมีค่าระดับเกณฑ์การแบ่งกลุ่มลดลง เมื่อระดับเกณฑ์เป็น 0.1 โมเดล SR ให้ค่า TP Rate สูงสุดคิดเป็น 0.6994 รองลงมาคือ โมเดล BNE และ BNLB ที่ให้ค่า TP Rate เท่ากันคือ 0.5318 ตามมาด้วยโมเดล BNLT, S และ DTL ตามลำดับ

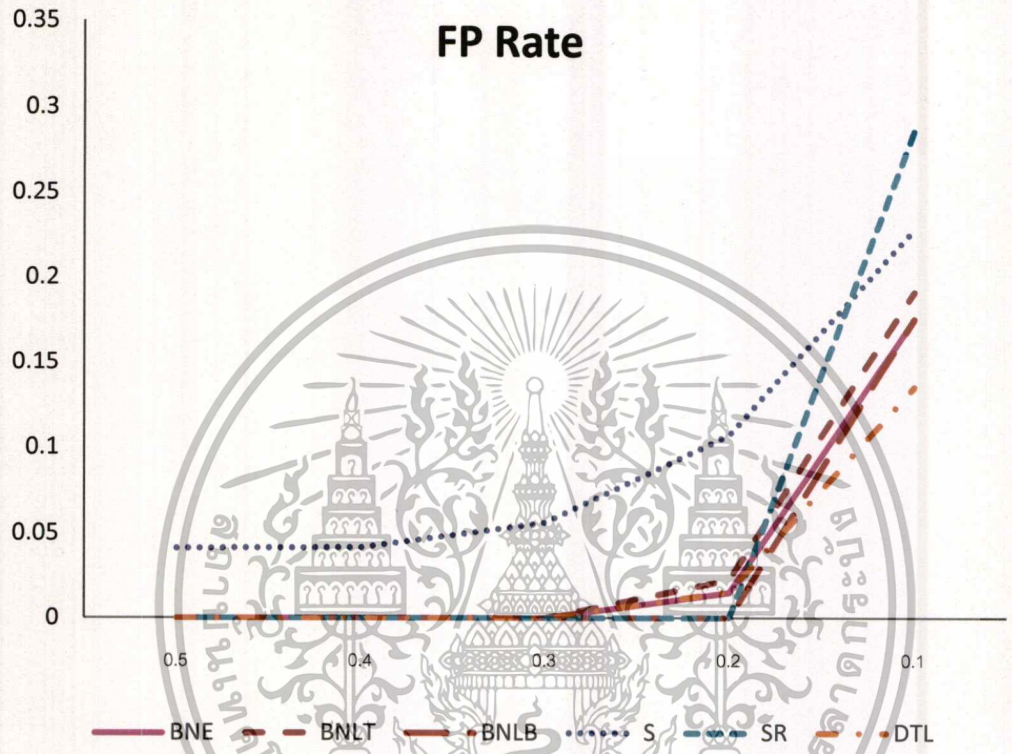


รูป 4-9 การเปรียบเทียบค่า TP Rate ที่ระดับเกณฑ์ในการแบ่งกลุ่ม (Threshold) ที่ต่างกัน



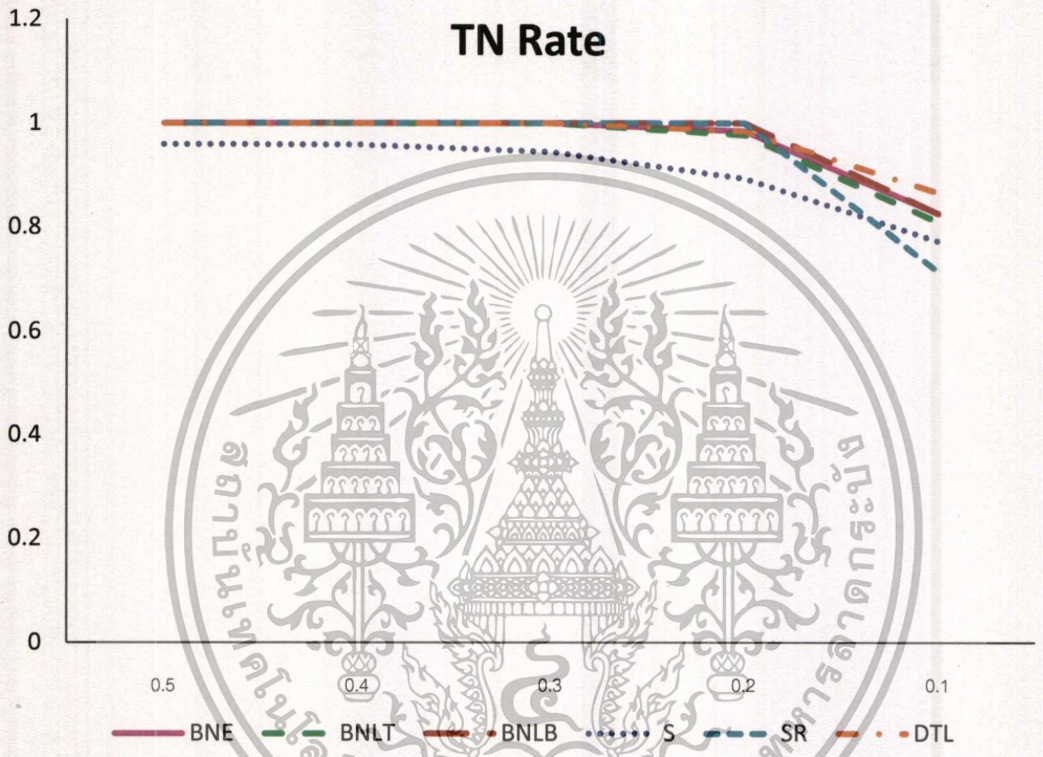
รูป 4-10 การเปรียบเทียบค่า Recall ที่ระดับเกณฑ์ในการแบ่งกลุ่ม (Threshold) ที่ต่างกัน

จากรูป 4-11 ทำให้ทราบว่าค่า FP Rate มีแนวโน้มเพิ่มขึ้นเมื่อมีค่าระดับเกณฑ์การแบ่งกลุ่มลดลง เมื่อระดับเกณฑ์เป็น 0.1 โมเดล DTL ให้ค่า FP Rate ต่ำสุดคิดเป็น 0.1353 รองลงมาคือโมเดล BNE และ BNLB ที่ให้ค่า FP Rate เท่ากันคือ 0.1754 ตามมาด้วยโมเดล BNLT, S และ SR ตามลำดับ



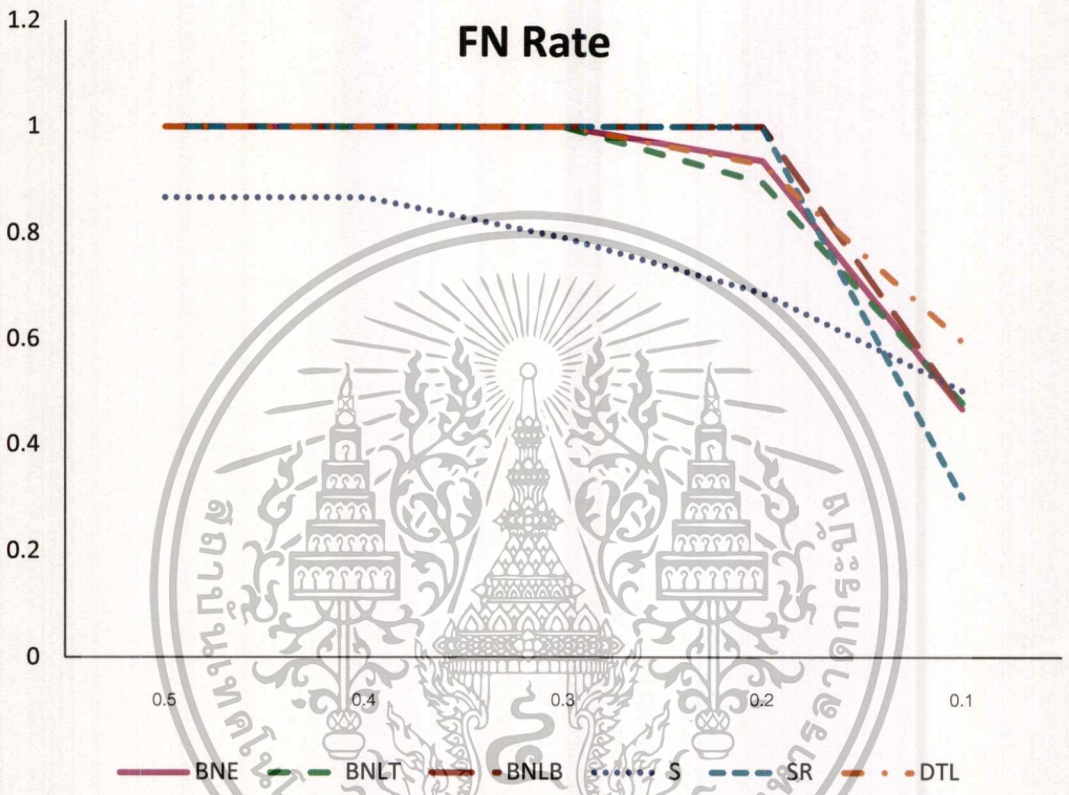
รูป 4-11 การเปรียบเทียบค่า FP Rate ที่ระดับเกณฑ์ในการแบ่งกลุ่ม (Threshold) ที่ต่างกัน

จากรูป 4-12 ทำให้ทราบว่าค่า TN Rate มีแนวโน้มลดลง เมื่อมีค่าระดับเกณฑ์การแบ่งกลุ่มลดลง เมื่อระดับเกณฑ์เป็น 0.1 โมเดล DTL ให้ค่า FP Rate สูงสุดคิดเป็น 0.8647 รองลงมาคือโมเดล BNE และ BNLB ที่ให้ค่า TN Rate เท่ากันคือ 0.8246 ตามมาด้วยโมเดล BNLT, S และ SR ตามลำดับ



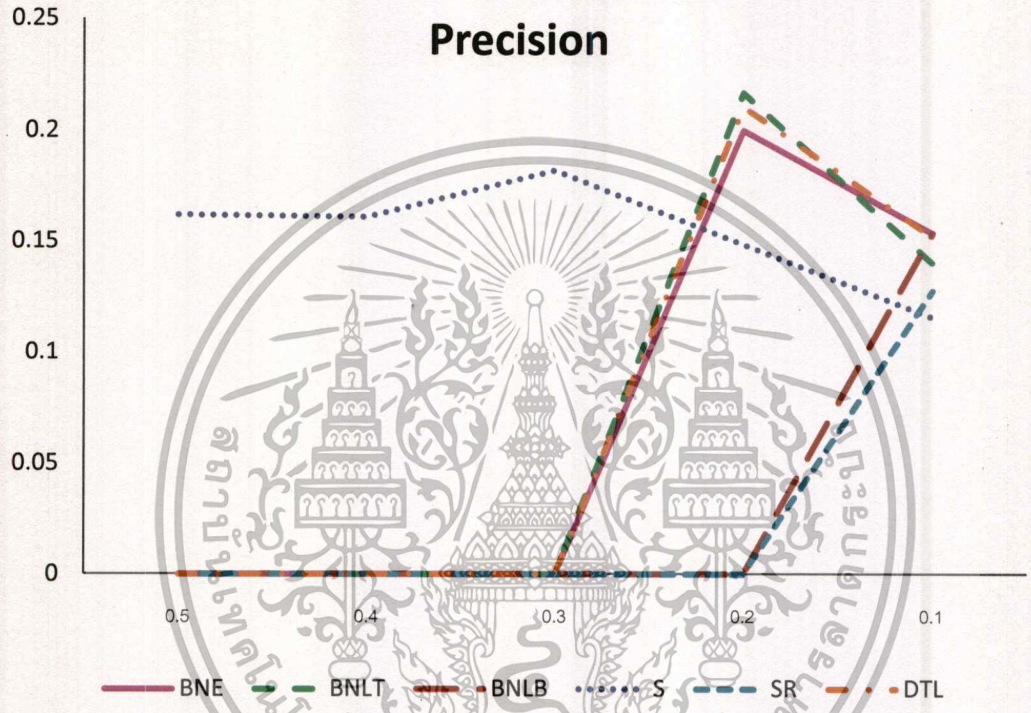
รูป 4-12 การเปรียบเทียบค่า TN Rate ที่ระดับเกณฑ์ในการแบ่งกลุ่ม (Threshold) ที่ต่างกัน

จากรูป 4-13 ทำให้ทราบว่าค่า FN Rate มีแนวโน้มลดลง เมื่อมีค่าระดับเกณฑ์การแบ่งกลุ่มลดลง เมื่อระดับเกณฑ์เป็น 0.1 โมเดล SR ให้ค่า FN Rate ต่ำสุดคิดเป็น 0.3006 รองลงมาคือ โมเดล BNE และ BNLB ที่ให้ค่า FN Rate เท่ากันคือ 0.4682 ตามมาด้วยโมเดล BNLT, S และ DTL ตามลำดับ



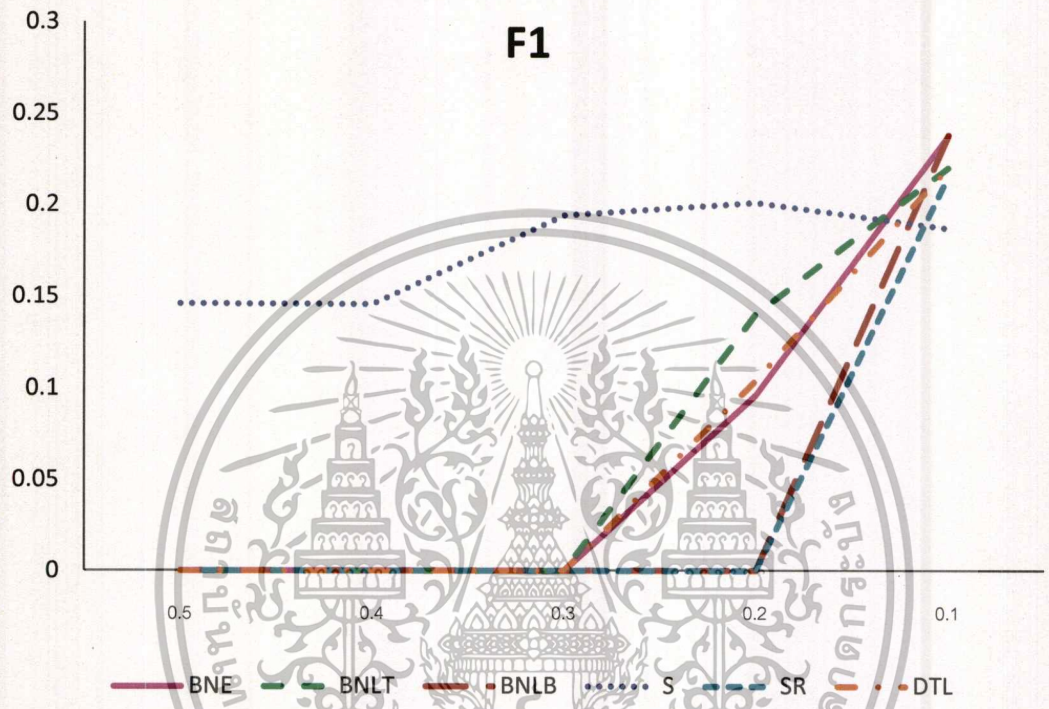
รูป 4-13 การเปรียบเทียบค่า FN Rate ที่ระดับเกณฑ์ในการแบ่งกลุ่ม (Threshold) ที่ต่างกัน

จากรูป 4-14 ทำให้ทราบว่าค่า Precision สามารถคำนวณได้เฉพาะช่วงค่าระดับแบ่งกลุ่มระหว่าง 0.1-0.2 โดยพบว่ามีแนวโน้มลดลงในช่วงดังกล่าว เมื่อค่าระดับเกณฑ์การแบ่งกลุ่มลดลง เมื่อระดับเกณฑ์เป็น 0.1 โมเดล BNE และ BNLB ให้ค่า Precision สูงสุดคิดเป็น 0.1536 รองลงมาคือโมเดล DTL ที่ให้ค่า Precision เท่ากับ 0.1518 อันดับ 3 คือ โมเดล BNLT ที่ให้ค่า Precision เท่ากับ 0.1400 ตามมาด้วยโมเดล SR และ S ตามลำดับ



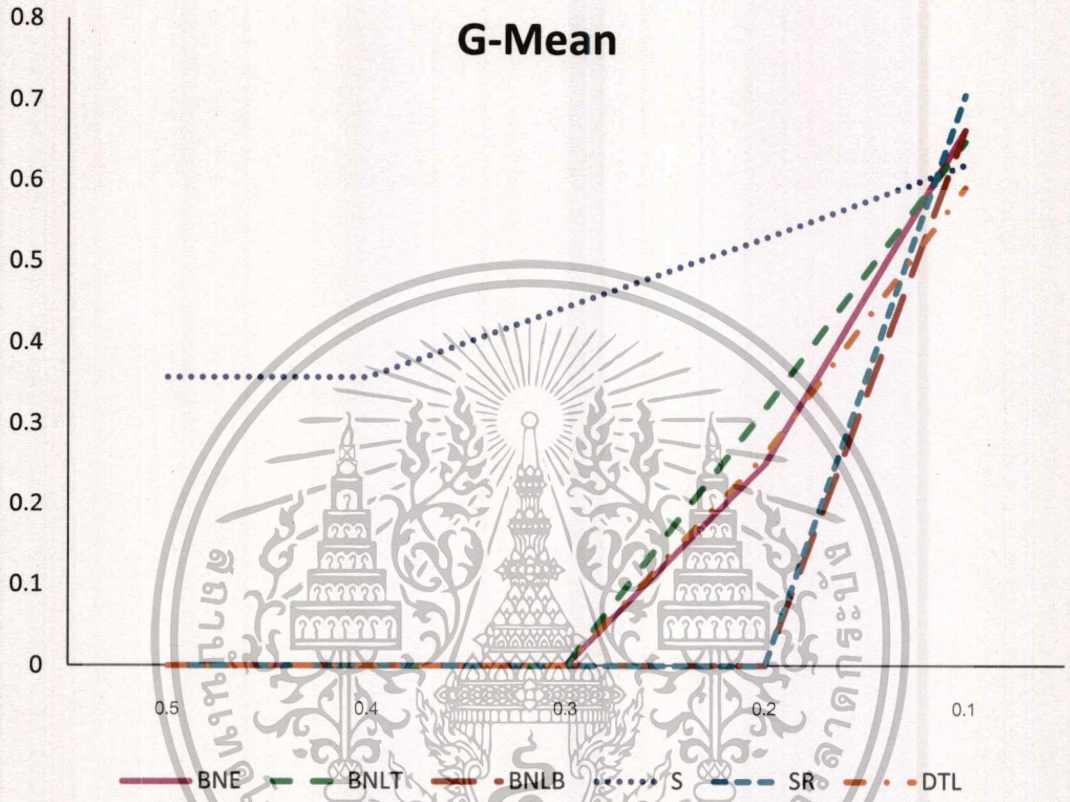
รูป 4-14 การเปรียบเทียบค่า Precision ที่ระดับเกณฑ์ในการแบ่งกลุ่ม (Threshold) ที่ต่างกัน

จากรูป 4-15 ทำให้ทราบว่าค่า F1 มีแนวโน้มเพิ่มขึ้นเมื่อมีค่าระดับเกณฑ์การแบ่งกลุ่มลดลง เมื่อระดับเกณฑ์เป็น 0.1 โมเดล BNE และ BNLB ให้ค่า F1 สูงสุดคิดเป็น 0.2383 รองลงมาคือ โมเดล DTL ที่ให้ค่า F1 เท่ากับ 0.2208 อันดับ 3 คือ โมเดล BNLT ที่ให้ค่า F1 เท่ากับ 0.2206 ตามมาด้วย โมเดล SR และ S ตามลำดับ



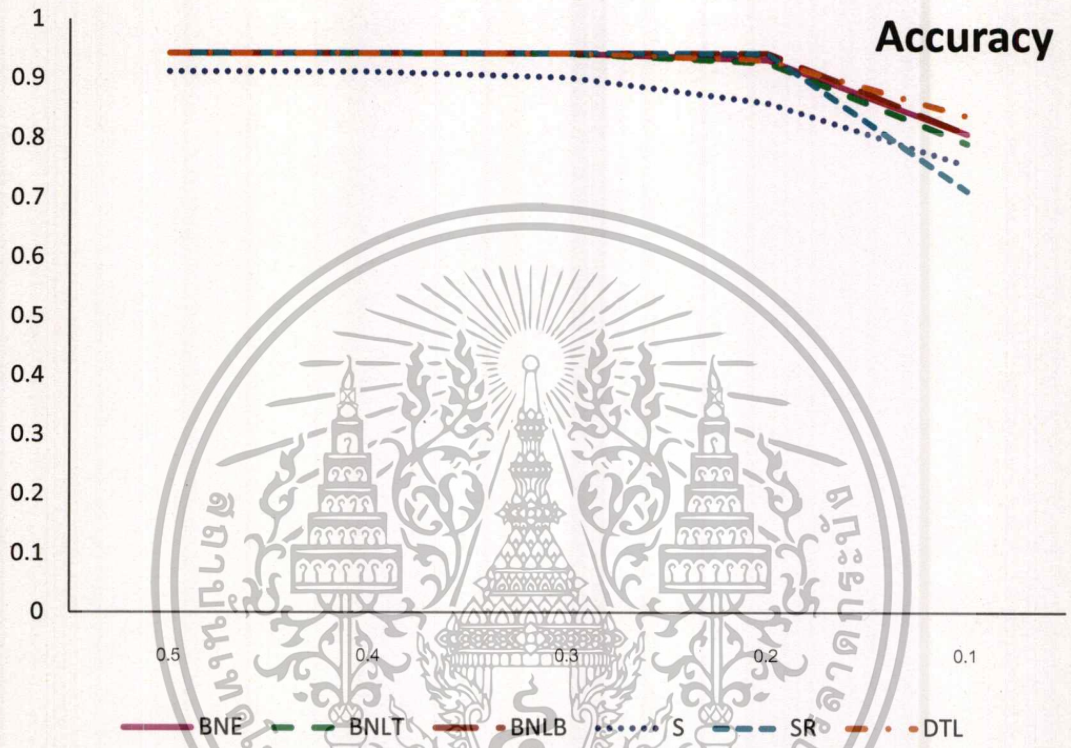
รูป 4-15 การเปรียบเทียบค่าคะแนน F1 ที่ระดับเกณฑ์ในการแบ่งกลุ่ม (Threshold) ที่ต่างกัน

จากรูป 4-16 ทำให้ทราบว่าค่า G-mean มีแนวโน้มเพิ่มขึ้นเมื่อมีค่าระดับเกณฑ์การแบ่งกลุ่มลดลง เมื่อระดับเกณฑ์เป็น 0.1 พบว่า โมเดล SR ให้ค่า G-mean สูงสุดคิดเป็น 0.7064 รองลงมาคือโมเดล BNE และ BNLB ให้ค่า G-mean เป็น 0.6622 อันดับ 3 ได้แก่ โมเดล BNLT ที่ให้ค่า G-mean เท่ากันคือ 0.6486 ตามมาด้วยโมเดล S และ DTL ตามลำดับ



รูป 4-16 การเปรียบเทียบค่าเฉลี่ยแบบเรขาคณิต (G-Mean) ที่ระดับเกณฑ์ในการแบ่งกลุ่ม (Threshold) ที่ต่างกัน

จากรูป 4-17 ทำให้ทราบว่าค่า Accuracy มีแนวโน้มลดลง เมื่อมีค่าระดับเกณฑ์การแบ่งกลุ่มลดลง เมื่อระดับเกณฑ์เป็น 0.1 พบว่า โมเดล DTL ให้ค่า Accuracy สูงสุดคิดเป็น 0.8387 รองลงมาคือโมเดล BNE และ BNLB ให้ค่า Accuracy เป็น 0.8080 อันดับ 3 ได้แก่ โมเดล BNLT ที่ให้ค่า Accuracy เท่ากันคือ 0.7924 ตามมาด้วยโมเดล S และ SR ตามลำดับ



รูป 4-17 การเปรียบเทียบค่าความถูกต้องในการจำแนก (G-Mean) ที่ระดับเกณฑ์ในการแบ่งกลุ่ม (Threshold) ที่ต่างกัน

บทที่ 5 สรุปผลการวิจัย

สำหรับงานวิจัยนี้มีวัตถุประสงค์เพื่อวิเคราะห์เปรียบเทียบโมเดลเครือข่ายแบบเบย์ที่ใช้ในการจำแนกกลุ่มผู้เป็นโรคเบาหวานที่สร้างขึ้น 6 โมเดล ซึ่งประกอบด้วย

1. BNE (BN_Expert) ใช้เทคนิค BN โดยมีลักษณะโครงสร้างแบบ Hierarchical และใช้วิธีการ Expert ในการกำหนดโครงสร้างและค่าความน่าจะเป็นจาก Training dataset

2. BNLB (BN_Learning with Bottom-up) ใช้เทคนิค BN โดยมีลักษณะโครงสร้างแบบ Hierarchical และใช้วิธีการ Learning (การเรียนรู้จากเครื่อง) ในการกำหนดโครงสร้างและค่าความน่าจะเป็นจาก Training dataset

3. BNLT (BN_Learning with Top-down) ใช้เทคนิค BN โดยมีลักษณะโครงสร้างแบบ Hierarchical และใช้วิธีการ Learning (การเรียนรู้จากเครื่อง) ในการกำหนดโครงสร้างและค่าความน่าจะเป็นจาก Training dataset

4. S (BN_Simple) ใช้เทคนิค BN โดยมีลักษณะโครงสร้างแบบ Non-Hierarchical และใช้วิธีการ Simple ในการกำหนดโครงสร้าง และค่าความน่าจะเป็นจาก Training dataset

5. SR (BN_SimReduce) ใช้เทคนิค BN โดยมีลักษณะโครงสร้างแบบ Non-Hierarchical และใช้วิธีการ Simple ในการกำหนดโครงสร้าง โดยเลือกเฉพาะตัวแปรที่สัมพันธ์กับการเกิดโรคเบาหวานอย่างมีนัยสำคัญ และกำหนดค่าความน่าจะเป็นจาก Training dataset

6. DTL (DT_Learning) ใช้เทคนิค Decision Tree โดยมีลักษณะโครงสร้างแบบ Hierarchical และใช้วิธีการ Learning (การเรียนรู้จากเครื่อง) ในการกำหนดโครงสร้างและค่าความน่าจะเป็นจาก Training dataset

ผลจากการเปรียบเทียบทั้ง 9 เกณฑ์ทำให้สามารถสรุปผลได้ดังต่อไปนี้

โมเดล BNE และ BNLB มีค่าความสามารถในการจำแนกที่ประเมินจาก Testing dataset ได้ดีกว่าโมเดลอื่น เพราะมีลำดับความสามารถในการจำแนกในแต่เกณฑ์ระหว่างลำดับที่ 1 หรือลำดับที่ 2 ซึ่งดีกว่าโมเดลอื่นๆ อย่างเห็นได้ชัด เมื่อพิจารณาผลการวิเคราะห์ด้วยค่า AUC พบว่าโมเดล BNE ดีกว่า BNLB เนื่องจากมีค่า AUC อยู่ในลำดับที่ 2 (สำหรับ Training dataset) และลำดับที่ 1 (สำหรับ Testing dataset) ดังนั้นจึงสรุปได้ว่า BNE เป็นโมเดลที่ดีกว่าโมเดลที่เหลืออีก 5 โมเดลที่เหลือ

ผลจากการเปรียบเทียบโมเดลที่ได้จากโมเดลเครือข่ายแบบเบย์ที่สร้างขึ้นโดยผู้เชี่ยวชาญ ซึ่งได้แก่ โมเดล BNE หรือที่สร้างขึ้นจากการเรียนรู้ด้วยเครื่องซึ่งได้แก่ โมเดล BNLT และ BNLB พบว่า ความสามารถในการจำแนกกลุ่มที่ได้จากคอนฟูชัน แมททริก ของ BNE และ BNLB ให้ แมททริก เหมือนกัน ซึ่งแต่แตกต่างจากคอนฟูชัน แมททริก ของ BNLT จึงทำ BNE และ BNLB มีความสามารถในการจำแนกกลุ่มได้ดีกว่า BNLT ซึ่งสอดคล้องกับค่า AUC ที่ทำให้ทราบว่า โมเดล BNLT มีค่า AUC อยู่ในลำดับที่ 5 ในขณะที่ BNE และ BNLB อยู่ในลำดับที่ 1 และ 2 ตามลำดับ ดังนั้นโดยสรุปแล้วจะเห็นได้ว่า การสร้างโมเดลที่ได้จากผู้เชี่ยวชาญมีประสิทธิภาพดีกว่าการสร้างโมเดลจากการเรียนรู้ด้วยเครื่องที่นำมาเปรียบเทียบในการศึกษานี้เล็กน้อยเมื่อเทียบกับโมเดล BNLB (ซึ่งพบว่ามีรูปแบบโมเดลที่ใกล้เคียงกับโมเดล BNE)

ผลจากการเปรียบเทียบระหว่างโมเดลเครือข่ายแบบเบย์ (BN) ที่มีโครงสร้างแบบลำดับขั้น (Hierarchical) ซึ่งได้แก่ โมเดล BNE BNLT และ BNLB และโมเดลแบบไม่เป็นลำดับขั้น (Non-Hierarchical) ซึ่งได้แก่ โมเดล S และ SR พบว่า โมเดลเครือข่ายแบบเบย์ (BN) ที่มีโครงสร้างแบบลำดับขั้น มีความสามารถในการพยากรณ์และการจำแนกกลุ่มที่ดีกว่าโมเดลแบบไม่เป็นลำดับขั้นโดยเฉพาะอย่างยิ่ง S ที่มีค่า AUC เพียง 0.6685 และหากทำการลดตัวแปรเพื่อสร้าง SR สามารถเพิ่มค่า AUC เป็น 0.7741 ซึ่งอยู่ในลำดับที่ 3 เมื่อเทียบกับโมเดลที่เหลือ และให้ค่า TP Rate และ Recall สูงกว่าโมเดลอื่นๆ ดังนั้นโดยสรุปแล้วจะเห็นได้ว่า โมเดลที่โครงสร้างแบบลำดับขั้น (Hierarchical) มีประสิทธิภาพดีกว่าโมเดลแบบไม่เป็นลำดับขั้น (Non-Hierarchical)

สุดท้ายผลการเปรียบเทียบของโมเดล BN กับ Decision Tree พบว่า โมเดล Decision Tree ที่สร้างขึ้นซึ่งคือโมเดล DTL ให้ค่า AUC อยู่ในลำดับที่ 4 นอกจากนี้ยังมีค่า Precision สูงแต่ Recall ต่ำ ทำให้เมื่อหาค่าเฉลี่ยของค่าทั้ง 2 ค่า แสดงด้วยค่า F1 มีค่าในระดับปานกลาง แต่เมื่อเทียบกันโมเดล BNE แล้วยังมีประสิทธิภาพในการจำแนกด้อยกว่า จึงสรุปผลจากโมเดลที่นำมาเปรียบเทียบว่าโมเดล BN ให้ค่าพยากรณ์ที่ดีกว่า Decision Tree

เอกสารอ้างอิง

อมรา ทองหงษ์, กมลชนก เทพสิทธา, & ภาคภูมิ จงพิริยะอนันต์. (2556). รายงานการเฝ้าระวังโรคไม่ติดต่อเรื้อรัง ปี พ.ศ. 2554. รายงานการเฝ้าระวังทางระบาดวิทยาประจำสัปดาห์, *44*(10), 145–152. Retrieved from http://www.boe-wesr.net/index.php

สำนักงานสำรวจสุขภาพประชาชนไทย. (2552). *รายงานการสำรวจสุขภาพประชาชนไทยโดยการตรวจร่างกาย: ครั้งที่ 4 พ.ศ. 2551-2552*. (เอกพลากรวิชัย, Ed.). กรุงเทพฯ: บริษัท เดอะ กราฟิโก ซิสเต็มส์ จำกัด. Retrieved from http://www.hiso.or.th/hiso/picture/reportHealth/report/report1.pdf

สำนักโรคไม่ติดต่อ กรมควบคุมโรค. (2556). *รายงานประจำปี พ.ศ. 2557*. กรุงเทพฯ. Retrieved from https://www.google.co.th/?gws_rd=cr&ei=GOgoVMqTIoqRuATuzoGYBg#q=รายงานการเฝาระวังโรคไม่ติดต่อเรื้อรัง+พ.ศ.+2557

Aekplakorn, W., Abbott-Klafter, J., Premgamone, A., Dhanamun, B., Chaikittiporn, C., Chongsuvivatwong, V., ... Lim, S. S. (2007). Prevalence and management of diabetes and associated risk factors by regions of Thailand: Third National Health Examination Survey 2004. *Diabetes Care*, *30*(8), 2007–12. http://doi.org/10.2337/dc06-2319

Atoui, H., Fayn, J., Gueyffier, F., & Rubel, P. (2006). Cardiovascular risk stratification in decision support systems: A probabilistic approach. application to pHealth.

Bouckaert, R. R. (2008). *Bayesian Network Classification in Weka for Version 3-5-7*. University of Wailato.

Burnside, E. S., Rubin, D. L., Fine, J. P., Shachter, R. D., Sisney, G. A., & Leung, W. K. (2006). Bayesian network to predict breast cancer risk of mammographic microcalcifications and reduce number of benign biopsy results: initial experience. *Radiology*, *240*(3), 666–73. http://doi.org/10.1148/radiol.2403051096

Cléret, M., Le Duff, F., Fresnel, A., & Le Beux, P. (2001). DIAMED: a probabilistic diagnostic aid system on the web. *Studies in Health Technology and Informatics*, *84*(Pt 1), 429–33. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11604776

Danaei, G., Finucane, M. M., Lu, Y., Singh, G. M., Cowan, M. J., Paciorek, C. J., ... Ezzati, M. (2011). National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2·7 million participants. *Lancet*, *378*(9785), 31–40. http://doi.org/10.1016/S0140-6736(11)60679-X

Duijm, N. J. (2009). Safety-barrier diagrams as a safety management tool. *Reliability Engineering & System Safety*, *94*(2), 332–341. http://doi.org/10.1016/j.ress.2008.03.031

Fellaji, S., Azmani, A., & Akharif, A. (2014). Bayesian approach for minimizing nephropathy risk for patients with type 2 diabetes. In *2014 9th International Conference on Intelligent Systems: Theories and Applications (SITA-14)* (pp. 1–4). IEEE. http://doi.org/10.1109/SITA.2014.6847311

Jensen, F. V, & Nielsen, T. D. (2007). *Bayesian networks and decision Graphs* (2nd ed.). New York: Springger.

Julia Flores, M., Nicholson, A. E., Brunskill, A., Korb, K. B., & Mascaro, S. (2011). Incorporating expert knowledge when learning Bayesian network structure: A medical case study. *Artificial Intelligence in Medicine*, *53*(3), 181–204. http://doi.org/10.1016/J.ARTMED.2011.08.004

Kjaerulff, U. B., & Madsen, A. L. (2008). *Bayesian networks and influence diagrams: a guide to construction and analysis*. New York: Springer.

Kudikyala, U. K., Bugudapu, M., & Jakkula, M. (2018). Graphical Structure of Bayesian Networks by Eliciting Mental Models of Experts (pp. 333–341). Springer, Singapore. http://doi.org/10.1007/978-981-10-5544-7_32

Kurt, I., Ture, M., & Kurum, A. T. (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Systems with Applications*, *34*(1), 366–374. http://doi.org/10.1016/j.eswa.2006.09.004

Lauritzen, S. L., & Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, *50*(2), 157–224. Retrieved from http://www.jstor.org/stable/2345762

Leerojanaprapa, K., Atthirawong, W., Aekplakorn, W., & Sirikasemsuk, K. (2017). Applying Bayesian Network for Noncommunicable Diseases Risk Analysis: Implementing National Health Examination Survey in Thailand. In *Industrial Engineering and Engineering Management (IEEM)*. Singapore.

Liu, K. F.-R., Lu, C.-F., Chen, C.-W., & Shen, Y.-S. (2011). Applying Bayesian belief networks to health risk assessment. *Stochastic Environmental Research and Risk Assessment*, *26*(3), 451–465. http://doi.org/10.1007/s00477-011-0470-z

Marcot, B. (2012). Metrics for evaluating performance and uncertainty of Bayesian network models. *Ecological Modelling*, *230*(null), 50–62. http://doi.org/10.1016/j.ecolmodel.2012.01.013

Sebastiani, P., Nolan, V. G., Baldwin, C. T., Abad-Grau, M. M., Wang, L., Adewoye, A. H., ... Steinberg, M. H. (2007). A network model to predict the risk of death in sickle cell disease. *Blood*, *110*(7), 2727–35. http://doi.org/10.1182/blood-2007-04-084921

Suvisaari, J., Loo, B.-M., Saarni, S. E., Haukka, J., Perälä, J., Saarni, S. I., ... Jula, A. (2011). Inflammation in psychotic disorders: a population-based study. *Psychiatry Research*, *189*(2), 305–11. http://doi.org/10.1016/j.psychres.2011.07.006

Ture, M., Kurt, I., Turhankurum, A., & Ozdamar, K. (2005). Comparing classification techniques for predicting essential hypertension. *Expert Systems with Applications*,

29(3), 583–588. http://doi.org/10.1016/j.eswa.2005.04.014

Twardy, C., Nicholson, A., & Korb, K. (2005). Knowledge engineering cardiovascular Bayesian networks from the literature. Retrieved from http://www.csse.monash.edu.au/publications/2005/tr-2005-170-full.pdf

Xiang, Z., Minter, R. M., Bi, X., Woolf, P. J., & He, Y. (2007). miniTUBA: medical inference by network integration of temporal data using Bayesian analysis. *Bioinformatics (Oxford, England)*, *23*(18), 2423–32. http://doi.org/10.1093/bioinformatics/btm372

Yang Guo, Guohua Bai, & Yan Hu. (2012). Using Bayes Network for Prediction of Type-2 diabetes.

เอกสารออนไลน์

ทีมข่าว MGR Online. (6 ก.พ. 2556). 'ไทยสู่ยุค "คนชราเต็มเมือง" จี้รัฐกันเงินก้อนโตไว้ดูแล หมอแนะผู้สูงวัยตุนเงิน "3 ล้านบาท" ไว้รักษา 2 โรคยอดฮิต!'. *Manager online*. สืบค้นเพื่อวันที่ 25 มีนาคม 2558, จาก **http://www.manager.co.th/Home/ViewNews.aspx?NewsID=9560000015532**

กิตติกรรมประกาศ

งานวิจัยนี้ได้รับการสนับสนุนงบวิจัยจากทุนอุดหนุนการวิจัย ประเภทเงินอุดหนุนทั่วไป (งบประมาณเงินรายได้) จากคณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ประจำปี 2559 นอกจากนี้ยังได้รับการสนับสนุนด้านข้อมูลจากสำนักงานการสำรวจสภาวะสุขภาพของประชาชนไทย (สสท.) ในการขออนุเคราะห์ข้อมูลจากฐานข้อมูลการสำรวจสุขภาพของประชาชนไทย โดยมีนายแพทย์ วิชัย เอกพลากร หัวหน้าภาควิชาเวชศาสตร์ชุมชน คณะแพทยศาสตร์โรงพยาบาลรามาธิบดี สนับสนุนและให้ความคิดเห็นด้านการสร้างโมเดล