ระบบทางการแพทย์อัจฉริยะสำหรับทำนายโรคเบาหวาน

ชนัดดาภรณ์ อ่อนแสง[1] และ ธนมณฑ์ ตันศิริเสริญกุล[2]

[1]คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง กรุงเทพฯ

[2]คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง กรุงเทพฯ

Emails: 58070025@it.kmitl.ac.th, 58070053@it.kmitl.ac.th

บทคัดย่อ

ในปัจจุบันผู้ป่วยจำนวนมากประสบกับโรคต่าง ๆ ที่ระบุความผิดปกติของร่างกายมนุษย์ ซึ่งต้องใช้กระบวนการที่มีความซับซ้อนจากห้องปฏิบัติการ คณะผู้จัดทำได้ยกประเด็นเรื่องโรคเบาหวาน ซึ่งเป็นหนึ่งในปัญหาสำคัญของประเทศตามที่รายงานไว้ในสถิติระดับประเทศต่าง ๆ นอกจากนี้โรคเบาหวานมักนำไปสู่การเกิดโรคเรื้อรังและเป็นอันตรายถึงชีวิตหากไม่ได้รับการตรวจพบแต่เนิ่น ๆ ดังนั้นการวินิจฉัยว่าเป็นโรคเบาหวานเร็วที่สุดเท่าที่จะเป็นไปได้ จึงเป็นสิ่งสำคัญอย่างยิ่งสำหรับผู้ป่วยที่จะได้รับการรักษาที่เหมาะสมเป็นการรักษาเชิงป้องกัน

โครงงานนี้เป็นการศึกษาและพัฒนาด้วยเทคนิคการเรียนรู้ด้วยเครื่องหลากหลายวิธี เพื่อทำนายโรคเบาหวานจากข้อมูลย้อนหลัง เรานำเสนอและทดลองตัวจำแนกของ Multi-Kernels on Multi-Layers โดยโมเดลที่ผ่านการฝึกอบรมนั้นถูกนำไปใช้กับเว็บแอปพลิเคชันเราพัฒนาโดย Django framework เพื่อความสะดวกในการใช้งานสำหรับแพทย์ พยาบาล และนักวิทยาการข้อมูล เว็บแอปพลิเคชันนี้เรียกว่าระบบการแพทย์อัจฉริยะสำหรับวินิจฉัยโรคเบาหวาน โดยข้อมูลที่ใช้ในโครงการนี้ได้รับการสนับสนุนจากสำนักโรคไม่ติดต่อ กรมควบคุมโรค กระทรวงสาธารณสุข (Thai NCD)

คำสำคัญ – การรวบรวมและจัดเตรียมข้อมูล; การประมวลผลข้อมูลขั้นต้น; การสร้างเวกเตอร์ของข้อมูล; การจัดหมวดหมู่และการจัดกลุ่ม; การแสดงแบบจำลองและการทำนาย

1. บทนำ

ตามรายงานสถิติโรคเบาหวานแห่งชาติปี 2015 มีผู้ป่วยโรคเบาหวาน จำนวน 30.3 ล้านคนในสหรัฐอเมริกา ซึ่งได้รับการวินิจฉัยเป็นโรคเบาหวานแล้ว 23.1 ล้านคน และยังไม่ได้รับการวินิจฉัยการเป็นโรคเบาหวานทั้งหมด 7.2 ล้านคน จาก 30.3 ล้านคน [1] สำหรับประเทศไทยจากรายงานของสำนักงานปลัดกระทรวงสาธารณสุข พบว่าอัตราการตายจากโรคเบาหวานต่อประชากรแสนคน ในภาพรวมของประเทศในปี 2556-2558 เท่ากับ 14.93, 17.53 และ 17.83 ตามลำดับ ซึ่งเพิ่มสูงขึ้นทุกปี และจากการสำรวจสุขภาพประชาชนไทยอายุ 15 ปีขึ้นไป โดยการตรวจสุขภาพของร่างกาย เมื่อปี 2557 พบว่า สถิติของการเป็นโรคเบาหวาน เพิ่มสูงขึ้นร้อยละ 8.9 คิดเป็นจำนวนมากถึง 4.8 ล้านคน เมื่อเทียบกับปี 2552 ซึ่งพบเพียงร้อยละ 6.9 หรือมีคนเป็นโรคเบาหวาน 3.3 ล้านคน [2] จากสถิติพบว่าภาวะแทรกซ้อนจากโรคเบาหวานส่งผลให้ผู้ป่วยต้องใช้เวลาในการเข้ารับการรักษาในโรงพยาบาลนานขึ้น และส่งผลกระทบต่อครอบครัวด้านค่าใช้จ่ายในการรักษา

การทำนายอาการโรคจำเป็นต่อการหาสาเหตุของโรค โดยการทำนายนั้นจะนำมาใช้ในเชิงป้องกัน เฝ้าระวัง ซึ่งโครงการนี้นำเสนอระบบทางการแพทย์อัจฉริยะสำหรับทำนายโรคเบาหวานได้ใช้เทคนิค Multi-Kernels on Multi-Layers ที่ทางคณะผู้จัดทำได้พัฒนาขึ้นในการสร้างโมเดล Machine Learning เพื่อทำนายโอกาสหรือแนวโน้มของผู้ป่วย

การวิจัยในการพัฒนาโมเดลและพัฒนาระบบทางการแพทย์นี้ เริ่มจากการขอชุดข้อมูลเบาหวานจากสำนักโรคไม่ติดต่อ กรมควบคุมโรค กระทรวงสาธารณสุข ซึ่งชุดข้อมูลผู้ป่วยประกอบด้วยส่วนหลัก ๆ คือ ข้อมูลส่วนตัวผู้ป่วยและข้อมูลประวัติการตรวจร่างกายต่าง ๆ ซึ่งจะนำข้อมูลเหล่านี้มาเรียนรู้เพื่อสร้างเป็นโมเดลที่ใช้ในการทำนาย และระบบทางการแพทย์อัจฉริยะฯ จะแสดงผลลัพธ์ในรูปแบบของเว็บแอปพลิเคชัน โดยจะแสดงรายละเอียดของข้อมูลผู้ป่วย ผลการทำนาย และแผนภูมิ

ภาพลักษณะต่าง ๆ ได้แก่ ร้อยละของผลการทำนายการเกิดโรคเบาหวานของผู้ป่วย และผลการตรวจประเมินสุขภาพ เป็นต้น

จากผลลัพธ์ที่แสดงในรูปแบบของเว็บแอป-พลิเคชันดังกล่าว แพทย์สามารถเลือกดูข้อมูลของผู้ป่วยเป็นรายบุคคลได้ เพื่อนำไปประกอบการตัดสินใจในการรักษาผู้ป่วย เลือกวิธีการรักษาที่เหมาะสมตามระดับอาการของโรค นอกจากนั้นการวินิจฉัยโรคเชิงทำนาย (Predictive Diagnosis) ยังมีประโยชน์ต่อทั้งผู้ป่วยและโรงพยาบาล เช่น แพทย์สามารถทำนายระยะเวลาในการเข้ารับการรักษาของผู้ป่วย การระบุหาตัวผู้ป่วยที่มีโอกาสเสี่ยงสูงต่อการเกิดโรคเบาหวานหรืออาการแทรกซ้อนต่าง ๆ ซึ่งเป็นส่วนสำคัญต่อแพทย์ในการเลือกวิธีดูแลและการรักษาอย่างทันท่วงทีต่อผู้ป่วยกลุ่มเสี่ยงเหล่านี้ อีกทั้งยังส่งเสริมการฟื้นหายและให้ผู้ป่วยมีประสบการณ์ที่ดีจากการเข้ารับการรักษา

2. การทบทวนวรรณกรรมที่เกี่ยวข้อง

2.1. ทฤษฎีที่เกี่ยวข้อง

การเรียนรู้ด้วยเครื่องจักร (Machine Learning) คือ เทคนิคหนึ่งที่สำคัญในด้านวิทยาการข้อมูล ซึ่งทำให้คอมพิวเตอร์สามารถเรียนรู้ได้ด้วยตนเอง โดยรูปแบบของการเรียนรู้ที่นำมาใช้คือ Supervised Learning เป็นการเรียนรู้ที่เน้นสอนคอมพิวเตอร์โดยการศึกษาจากข้อมูลตัวอย่าง (Label) ซึ่งสามารถแบ่งประเภทอัลกอริทึมออกเป็น 2 กลุ่ม ได้แก่ การจัดหมวดหมู่ของข้อมูล (Classification) และการถดถอย (Regression) โดยโมเดลหรือแบบจำลองการเรียนรู้ มี 2 แบบ ได้แก่

2.1.1 การเรียนรู้แบบเดี่ยว (Individual standard model)

แบบจำลองการเรียนรู้จากอัลกอริทึมเดียวไม่ใช้ควบคู่กับตัวอื่น โดยไม่ใช้ควบคู่กับตัวอื่น ซึ่งในงานวิจัยนี้ใช้ทั้งหมด 4 อัลกอริทึม ได้แก่

1) Decision tree [3] เป็นแบบจำลองทางคณิตศาสตร์ที่เรียนรู้ข้อมูลเพื่อหาทางเลือกที่ดีที่สุดโดยอยู่ในรูปแบบเงื่อนไข และผลลัพธ์ ซึ่งเงื่อนไขดังกล่าวจะแบ่งข้อมูลตามคุณลักษณะ (Feature) เพื่อนำไปสร้างโมเดลในการทำนาย

2) K-Nearest Neighbors [4] เป็นวิธีการค้นหาเพื่อนบ้านที่ใกล้ที่สุด หรือข้อมูลที่มีความคล้ายคลึงกันมากที่สุดเป็นจำนวน k ตัว โดยใช้การคำนวณระยะห่างระหว่างข้อมูลกลุ่มตัวอย่างกับข้อมูลที่กำลังสนใจ ซึ่งการคำนวณวิธีต่าง ๆ ที่ขึ้นอยู่กับประเภทของคุณลักษณะข้อมูลและการนำไปใช้

3) Logistic Regression [5] เป็นการประมาณการประเภทของข้อมูล โดยใช้ Logistic function สร้าง S-shaped curve ลากตัดผ่านกลุ่มข้อมูลตัวอย่างเพื่อทำนายว่าเป็นข้อมูลประเภทไหน ซึ่งเหมาะกับการทำนายข้อมูลประเภทจัดกลุ่ม

4) Support Vector Machine [4] เป็นการจัดหมวดหมู่เชิงเรขาคณิตซึ่งคำนวณหาไฮเปอร์เพลน หรือระนาบที่มีซับซ้อน เพื่อใช้ในการแบ่งหมวดหมู่ของข้อมูลในเวกเตอร์สเปซหลายมิติ โดยไฮเปอร์เพลนที่ดีที่สุด คือ ไฮเปอร์เพลนที่สามารถแบ่งข้อมูลออกห่างจากกันได้มากที่สุดเมื่อเทียบระยะห่างจากไฮเปอร์เพลน

2.1.2 เรียนรู้แบบกลุ่ม (Ensemble model)

การใช้หลายอัลกอริทึมมาช่วยในสร้างแบบจำลองการเรียนรู้ ซึ่งเป็นเทคนิคสร้างโมเดลด้วยการรวมผลลัพธ์การทำนายของแบบจำลองการเรียนรู้แบบเดี่ยวที่ได้ผลลัพธ์แตกต่างกันออกไปมาใช้ร่วมกันเพื่อเพิ่มประสิทธิภาพของโมเดล โดยที่ Ensembles มีหลักการทำงานพื้นฐาน แบ่งออกเป็นได้ 3 รูปแบบ ได้แก่ Bagging, Boosting และ Stacking [6] โดยในงานวิจัยนี้ได้เลือก Random Forest เป็นตัวอย่างของโมเดลการเรียนรู้แบบกลุ่มเพื่อมาเปรียบเทียบกับ Multi-Kernels on Multi-Layers

1) Random Forest เป็นโมเดลการเรียนรู้แบบกลุ่มที่ใช้เทคนิค Bagging ที่สุ่มสร้างข้อมูลทดสอบ (Training sets) ที่แตกต่างกัน ซึ่งผลลัพธ์คือการสร้างโมเดล Decision Tree ขึ้นมาหลายโมเดล และนำผลการทำนายมารวมกันโดยวิธีการโหวต (Voting)

2) Multi-Kernels on Multi-Layers เป็นโมเดลหลักที่นำเสนอในงานวิจัยนี้ ซึ่งใช้เทคนิครูปแบบ Stacking ที่เป็นการใช้การเรียนรู้แบบกลุ่ม โดยรวมอัลกอริทึมการทำนายที่แตกต่างมากกว่า 1 ประเภทมารวมไว้ในที่เดียว และมีการขั้นตอนการทำนายมากกว่า 1 ชั้น และมีการใช้ K-fold Cross validation เพื่อสร้างการรวมการถ่วง

น้ำหนักที่ดีที่สุดของการทำนาย อีกทั้งยังสามารถหลีกเลี่ยงการเกิด Overfitting

2.2. เทคนิคหรือเทคโนโลยีที่ใช้

2.2.1. Pandas

เป็นคลังฟังก์ชัน (Library) ในภาษา Python ที่ได้รับความนิยม เนื่องจากมีโครงสร้างข้อมูลที่ดีและเครื่องมือในการวิเคราะห์ข้อมูลครบ ทำให้จัดการกับข้อมูลได้ง่ายยิ่งขึ้นเหมาะในการทำ Data Cleaning [7]

2.2.2. NumPy

เป็นคลังฟังก์ชัน (Library) ในภาษา Python ซึ่งมีฟังก์ชันเกี่ยวกับคณิตศาสตร์ที่มีประสิทธิภาพสูง รวดเร็ว และมีการทำงานต่าง ๆ [8]

2.2.3. Matplotlib

เป็นไลบรารี่สำหรับการแสดงผลข้อมูล โดยการพลอตกราฟอาเรย์สองมิติ สามารถแสดงผลข้อมูลได้อย่างรวดเร็ว และบันทึกผลที่ได้ออกมาเป็นรูปภาพได้ [8]

2.2.4. Seaborn

ไลบรารี่ที่ถูกสร้างขึ้นบนพื้นฐานต่อจาก Matplotlib มีจุดเด่นในการแสดงผลในลักษณะของการสรุปข้อมูล

2.2.5. Scikit-learn

ไลบรารี่ที่นิยมในการทำ Machine Learning มีฟังก์ชันสำหรับการวัดประสิทธิภาพต่าง ๆ และครอบคลุมอัลกอริทึมของ Machine Learning แทบทั้งหมด [9]

2.3. เครื่องมือที่เกี่ยวข้อง

2.3.1. Google Colaboratory

เครื่องมือสำหรับทำโครงการวิจัย โดยมีลักษณะการทำงานแบบเดียวกับ Jupyter Notebook ซึ่งเป็นเว็บแอปพลิเคชัน และสามารถดำเนินงานทั้งหมดได้ในระบบคลาวด์ และแบ่งปันไฟล์โค้ดร่วมกับผู้อื่นได้

2.3.2. Python

ชื่อภาษาที่ใช้ในการเขียนโปรแกรม เหมาะสำหรับการทำวิทยาการคอมพิวเตอร์ เนื่องจากมี Pandas และ Scikit-learn เป็นคลังฟังก์ชัน (Library)

2.3.3. Django Framework

ชุดเครื่องมือสำหรับพัฒนาเว็บไซต์ด้วยภาษา Python ซึ่งมีคุณสมบัติ เช่น ORM เป็นต้น

2.4. งานวิจัยที่เกี่ยวข้อง

2.4.1. Zhou et al.

ได้เลือกใช้เทคนิค Deep Learning Feature Selection ในการคัดเลือกข้อมูลตัวแทนที่มีความ Compact representation โดยใช้ข้อมูลของผู้ป่วยโรคปอดบวม (Pneumonia) ในการศึกษา [10]

2.4.2. Cole and King

ใช้ข้อมูลยีน (gene) ในการทำนายโรคของผู้ป่วยติดเชื้อเอชไอวีโดยใช้เทคนิคการจัดหมวดหมู่ จากการศึกษาพบว่า rev gene มีค่าประสิทธิภาพของการทำนายระดับของโรคมากที่สุด [11]

2.4.3. Lippmann et al.

ใช้เทคนิค Neural Network ในการทำนายโอกาสการเสียชีวิต โรคหลอดเลือดสมอง (Stroke) และอาการไตวายของผู้ป่วยที่เข้ารับการผ่าตัดทำทางเบี่ยงหลอดเลือดหัวใจ จากการศึกษาพบว่า Multilayer perceptron โดยใช้ sigmoid เป็นฟังก์ชันการกระตุ้นของชั้นซ่อนตัว ที่มีประสิทธิภาพเหนือกว่าการใช้วิธีการถดถอยแบบ Logistic regression [12]

2.4.4. Wu and Wang

ได้ทำนายหาสาเหตุการเสียชีวิตเพื่อสุขภาพของประชากร โดยใช้ Deep Learning ได้ผลลัพธ์ค่าความถูกต้องประมาณ 75% [13]

3. วิธีการดำเนินการวิจัย

3.1 ปัญหาที่พบในปัจจุบัน

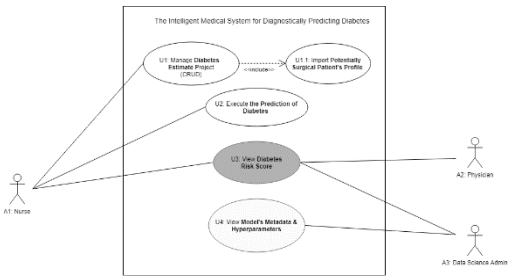
การเป็นโรคเบาหวานหากไม่ได้รับการตรวจพบหรือเฝ้าระวัง อาจนำไปสู่การแพร่กระจายก่อให้เกิดโรคต่าง ๆ หรืออันตรายถึงชีวิตได้ ดังเช่นกรณีตัวอย่างของ ตำรวจวัย 51 ปี เป็นโรคเบาหวาน ในขณะปฏิบัติหน้าที่เกิดอาการน้ำตาลในร่างกายเพิ่มสูงขึ้นมาก ทำให้เป็นลมหมดสติและเสียชีวิตในที่สุด [14]

จากกรณีดังกล่าวส่งผลกระทบต่อผู้ป่วยทั้งสุขภาพกาย สุขภาพจิตและค่าใช้จ่ายในการรักษา รวมไปถึงชีวิตของผู้ป่วย ทีมพัฒนาจึงได้เล็งเห็นว่าการสร้าง

แบบจำลองและการทำนายที่แสดงในรูปแบบเว็บไซต์ทำให้แพทย์สามารถดูผลจากการทำนายโอกาสการเกิดโรคเบาหวาน เพื่อนำไปประกอบการรักษาได้ง่ายและรวดเร็วยิ่งขึ้น

3.2 การวิเคราะห์และออกแบบระบบใหม่

3.2.1. Use Case Diagram



รูปที่ 1. แผนภาพยูสเคสของระบบ

รายละเอียดของแผนภาพยูสเคสมีดังนี้

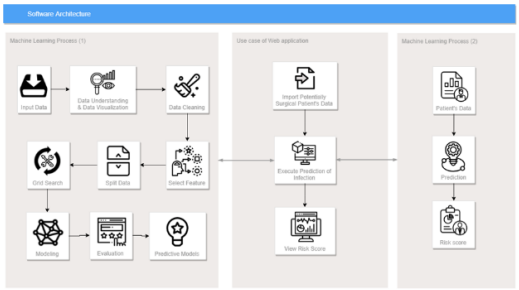
1. การจัดการทำนายโรคเบาหวาน โดยพยาบาลสามารถเพิ่ม, ลบ, ปรับปรุงแก้ไข และเรียกดูข้อมูลได้ ซึ่งมีกระบวนการนำข้อมูลผู้ป่วยเข้าสู่ระบบ

2. การดำเนินการทำนายโรคเบาหวาน พยาบาลจะนำข้อมูลผู้ป่วยเข้าสู่กระบวนการทำนายจากข้อมูลของผู้ป่วยที่เข้ารับการรักษา

3. ดูคะแนนโอกาสเสี่ยงการเป็นโรคเบาหวาน โดยพยาบาล แพทย์ และนักวิทยาศาสตร์ ซึ่งจะแสดงออกมาในรูปแบบแผนภูมิภาพ และข้อความตัวอักษร

4. ดูรายละเอียดที่อธิบายถึงความเป็นมาของโมเดล นักวิทยาศาสตร์ข้อมูล สามารถดูรายละเอียดเพิ่มเติมที่อธิบายถึงความเป็นมาของแต่ละโมเดล เช่น Accuracy score เป็นต้น เพื่อใช้ในการพัฒนาโมเดลต่อไปในอนาคต

3.2.2. โครงสร้างของซอฟต์แวร์



รูปที่ 2. โครงสร้างซอฟต์แวร์ของระบบ

ขั้นตอนการทำ Machine Learning (1)

1) Input Data คือ การนำข้อมูลของผู้ป่วยโรคเบาหวาน

2) Data Understanding and Visualization คือ ทำความเข้าใจรูปแบบและลักษณะของข้อมูลที่ได้มา โดยนำมาแสดงผลในรูปแบบของกราฟแบบต่าง ๆ

3) Data Cleaning คือ การตรวจสอบและแก้ไขชุดข้อมูล โดยมีการปรับปรุง หรือลบข้อมูลที่ไม่ถูกต้องออกไปจากชุดข้อมูล

4) Select Feature คือ การคัดเลือกปัจจัยสำคัญ (Feature) ที่มีความสัมพันธ์กับ Label

5) Split Data คือ การแบ่งข้อมูลเป็นทั้งหมด 2 ชุดข้อมูล ได้แก่ ข้อมูลฝึก 70% และข้อมูลทดสอบ 30% และทำการแบ่งเป็น 10-folds Cross Validation เพื่อใช้สร้างโมเดลกับจำนวนข้อมูลทั้งหมด

6) Grid Search คือ หาค่า parameter ที่ดีที่สุดของโมเดล ซึ่งผู้พัฒนาสามารถกำหนดค่า parameters

7) Modeling คือ การสร้างโมเดลจากปัจจัยสำคัญของข้อมูลโดยใช้อัลกอริทึมต่าง ๆ

8) Evaluation คือ การประเมินผล เพื่อเป็นการตัดสินคุณภาพของโมเดลโดยยึดค่าความถูกต้อง (Accuracy)

9) Predictive Models คือ โมเดลของของการทำนายการติดเชื้อและโมเดลของการทำนายโรคเบาหวาน ซึ่งโมเดลเหล่านี้จะถูกนำไปแสดงในเว็บแอปพลิเคชัน

ขั้นตอน Use case of Web application

1) Import Patient's Profile คือการนำข้อมูลผู้ป่วยเข้าสู่ระบบ

2) Execute Prediction of Infection คือ การนำข้อมูลและโมเดล

3) View Risk Score คือการดูคะแนนความเสี่ยงในการติดเชื้อจากการผ่าตัดหรือโรคเบาหวานของผู้ป่วยจะแสดงออกมาเป็นรูปแบบแผนภูมิ และข้อความ โดยเรียกข้อมูลมาจากผลลัพธ์ของขั้นตอนการทำ Machine Learning (2)

ขั้นตอนการทำ Machine Learning (2)

1) Patient’s Data คือ การเรียกข้อมูลผู้ป่วยที่ถูกเลือกเข้าสู่กระบวนการทำนาย

2) Prediction คือ การทำนายโดยใช้โมเดล และข้อมูลผู้ป่วยที่ถูกเลือก

3) Risk score คือ คะแนนความเสี่ยงการเป็นโรคเบาหวานหรือการติดเชื้อจากการผ่าตัดของผู้ป่วย

4. ผลการดำเนินงาน

4.1 การทำ Machine Learning

ผู้พัฒนาได้ใช้ชุดข้อมูลจากสำนักโรคไม่ติดต่อมาใช้ในการสร้างและพัฒนาโมเดลโดยต่อจากชุดข้อมูลโรคเบาหวานจากเว็บไซต์สถาบันโรคเบาหวาน โรคทางเดินอาหารและโรคไตแห่งชาติของประเทศสหรัฐอเมริกา เพื่อให้มีความเหมาะสมกับชาวไทยมากขึ้น ซึ่งชุดเบาหวาน ประกอบด้วยข้อมูล 2 กลุ่ม คือ กลุ่มผู้ที่ไม่เป็นเบาหวาน และเป็นโรคเบาหวาน ซึ่งมีขั้นตอน ดังนี้

4.1.1 Data Understanding

เป็นการทำความเข้าใจชุดข้อมูล ซึ่งชุดข้อมูลเบาหวานชุดนี้ ผู้ที่เข้ารับการตรวจเป็นคนไทย เพศชายและหญิง ประกอบด้วยข้อมูล 2 กลุ่มคือ กลุ่มผู้ที่ไม่เป็นเบาหวาน 4,234 คน และกลุ่มผู้ที่เป็นโรคเบาหวาน 4,171 คน รวมทั้งหมด 8,405 คน และมีรายละเอียดลักษณะของปัจจัยดังตารางที่ 1 นี้

ตารางที่ 1. พจนานุกรมชุดข้อมูลเบาหวาน

ชื่อ	คำอธิบาย
SEX	เพศ
AGE	อายุ
SMOKE	ประวัติการสูบบุหรี่
ALCOHOL	ประวัติดื่มเครื่องดื่มแอลกอฮอล์
DMFAMILY	ประวัติเบาหวานในญาติสายตรง
HTFAMILY	ประวัติความดันสูงในญาติ
WEIGHT	น้ำหนัก
HEIGHT	ส่วนสูง
WAIST_CM	เส้นรอบเอว
SBP_1	ซิสโตลิก วัดครั้งที่ 1
DBP_1	ไดแอสโตลิก วัดครั้งที่ 1
SBP_2	ซิสโตลิก วัดครั้งที่ 2
DBP_2	ไดแอสโตลิก วัดครั้งที่ 2

4.1.2 Data Cleaning

ผู้พัฒนาเริ่มจากการตรวจจับข้อมูลที่มีความผิดเพี้ยนไปจากค่าปกติ (Outlier) โดยการใช้ Interquartile range (IQR) เริ่มจากการหาค่า NaN หรือข้อมูลที่ไม่น่าจะมีเป็นค่า 0 ตามมาตรฐานตามความเป็นจริง ซึ่งค่า 0 สามารถหาได้จากกราฟ Box Plot จากนั้นจึงทำการแก้ไขข้อมูลที่ผิดพลาด มีวิธีคือการแทนที่ค่านั้นด้วยค่าเฉลี่ยของข้อมูลแต่ละฟีเจอร์ แต่จากการศึกษาเพิ่มเติมได้พบว่ามีอีกหลากหลายวิธีที่สามารถแทนค่าผิดพลาดดังกล่าว เช่น การทำนายค่าที่ผิดพลาดด้วย Machine Learning เป็นต้น หลังจากได้ทดลองหลาย ๆ วิธีผู้พัฒนาจึงเลือกที่จะแทนค่าผิดปกติด้วยค่าเฉลี่ยเนื่องจากมีประสิทธิภาพสูงสุดในการจัดการข้อมูลที่ผิดปกติของชุดข้อมูลนี้

4.1.3 Data Normalization

เป็นการทำข้อมูลให้เป็นรูปแบบ ซึ่งการทำ Normalize จะลดขนาดของข้อมูลให้มีขนาดเท่ากัน

4.1.4 Selecting Features

ใช้วิธีการคัดเลือกฟีเจอร์โดยการคำนวณ Pearson Correlation เพราะเหมาะกับการหาค่าสหสัมพันธ์ที่เป็นรูปแบบตัวเลขที่เป็นลักษณะข้อมูลเชิงปริมาณแบบต่อเนื่อง (Continues Data) ข้อมูลระดับช่วงชั้น (Interval Scale) และข้อมูลระดับอัตราส่วน (Ratio Scale) จากชุดข้อมูลเบาหวานผู้พัฒนาได้เลือก BMI, SBP_1, WAIST_CM, และ DBP_1 โดยมีค่า 0.26, 0.21, 0.29 และ 0.17 ตามลำดับดับ มาเป็นฟีเจอร์ในการทำนายเพราะมีค่าใกล้เคียง 1 หรือมีความสัมพันธ์กับ Label มากที่สุด เมื่อเทียบกับฟีเจอร์อื่น ๆ

4.1.5 Split Data and Grid SearchCV

เบื้องต้นจะทดสอบโมเดลด้วยการแบ่งข้อมูลออกเป็น 2 ชุด และได้ทำการหาค่า Hyperparameter ที่ดีที่สุดของโมเดล (Grid searchCV) เหมือนกับการทดลองรันโมเดล โดยใช้ข้อมูลในส่วนของ Training Data ผู้พัฒนาได้เลือกใช้อัลกอริทึมต่าง ๆ เช่น Multi-Kernels on Multi-Layers, Logistic Regression, Random Forest เป็นต้น และจากนั้นจึงทำการแบ่งชุดข้อมูลใหม่เป็น 10 ชุด ด้วย 10-folds Cross validation เพื่อสร้างโมเดลจริง

4.1.6 Modeling

ผู้พัฒนาได้มีการทำการทดลองทั้งแบบ Individual Standard Models และ Multi-Kernels on Multi-Layers โดยได้นำได้นำค่า Hyperparameter จากการทำ Grid SerachCV มาปรับค่าที่นำใส่ในโมเดล โดยมีอัลกอริทึมที่ใช้ในการเปรียบเทียบประสิทธิภาพของการทำนาย ได้แก่ Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT) และ K-Nearest Neighbor (KNN) ในส่วนของการทดลองแบบ Multi-Kernels on Multi-Layers จะแบ่งการฝึกข้อมูลออกเป็น 2 ชั้น ในชั้นแรกมีการกำหนดทั้งหมด 6 อัลกอริทึมได้แก่ DT, LR, KNN, Logistic RegressionCV, RF และ SVM และในชั้นที่สองจะกำหนดได้เพียง 1 อัลกอริทึมเท่านั้นจึงได้ทำการทดลอง 4 อัลกอริทึมเพื่อทำการคัดเลือกโมเดลที่ดีที่สุด ได้แก่ DT, SVM, LR และ RF โดยที่ Input ของชั้นแรกคือ Descriptive feature จากนั้นนำ Output ของชั้นแรกมาเป็น Input ของชั้นที่สอง และ Output ของชั้นที่สองคือผลลัพธ์การทำนายของ Multi-Kernels on Multi-Layers ซึ่งโมเดลจะได้รับการปรับแต่งคัดเลือกพารามิเตอร์ที่ทำให้ Multi-Kernels on Multi-Layers มีประสิทธิภาพสูงสุด ประกอบไปด้วย ชั้นที่หนึ่ง คือทั้ง 6 อัลกอริทึมเหมือนเดิม และชั้นที่สองคือ Logistic Regression

4.1.7 Evaluation

ในการวัดประสิทธิภาพทั้งแบบ Individual Standard Models และ Multi-Kernels on Multi-Layers ได้ใช้การเปรียบเทียบค่าความถูกต้อง (Accuracy) ค่าเอนเอียง (Bias) เป็นการวัดค่าเฉลี่ยของค่าที่ทำนายออกมาได้ ยิ่งค่าสูงหมายความว่าค่าทำนายไม่ใกล้เคียงกับค่าความจริง ค่าแปรปรวน (Variance) เป็นการวัดค่าความผิดพลาด (Error) ของความสามารถในการทำนายโมเดล หากมีค่าต่ำจะไม่เกิดปัญหาการเกิดอาการโมเดลที่ใช้ในการทำนายได้ดีกับข้อมูลฝึกมากเกินไป (Overfitting) และค่าความแม่นยำของโมเดล ได้ผลลัพธ์ดังตารางที่ 2

ตารางที่ 2. ผลการเปรียบเทียบประสิทธิภาพความถูกต้อง ค่าเอนเอียง ค่าแปรปรวน และค่าความแม่นยำของโมเดลทำนายโรคเบาหวาน ในชุดข้อมูลฝึก และทดสอบ

Algorithm	Individual Standard Model				Ensemble Model
	KNN	SVM	LR	RF	Multi-Kernels on Multi-Layers
Training Accuracy	0.77	0.88	0.66	0.83	0.78
Test Accuracy	0.70	0.69	0.66	0.71	**0.74**
Bias	0.23	0.12	0.34	0.12	0.22
Variance	0.07	0.19	0.00	0.12	0.04
Precision	0.69	0.70	0.66	0.68	0.73

จากตารางการเปรียบเทียบประสิทธิภาพในการทำนายของโมเดลทั้ง 4 รูปแบบ พบว่าเทคนิค Multi-Kernels on Multi-Layers นั้นให้ประสิทธิภาพที่ดีกว่าการเรียนรู้แบบ Individual Standard Model ทั้ง 5 รูปแบบ โดยมี Test Accuracy อยู่ที่ 0.74

5. บทสรุป

ระบบทางการแพทย์อัจฉริยะสำหรับทำนายโรคเบาหวาน เป็นเครื่องมือช่วยในการประเมินโรคเบาหวานในรูปแบบเว็บแอปพลิเคชัน เพื่ออำนวยความสะดวกในการใช้งานสำหรับแพทย์ พยาบาล และนักวิทยาการข้อมูล โดยนำเทคนิค Multi-Kernels on Multi-Layers (MKML) หรือ Stacking Ensembles หนึ่งในแขนงของเทคนิคการเรียนรู้ด้วยเครื่องจักร (Machine Learning) มาใช้ในการสร้างโมเดลทำนายโรคเบาหวานจากการทบทวนแผนภูมิย้อนหลังของผู้ป่วย โดยโครงงานนี้ได้เริ่มพัฒนาจากการศึกษาและทดลองสร้างโมเดลด้วยเทคนิค Machine Learning หลากหลายรูปแบบ คือ แบบเดี่ยว (Individual Standard Model) และแบบกลุ่ม (Ensemble Models) รวมถึงการปรับ

ค่าพารามิเตอร์และทำการเปรียบเทียบผลลัพธ์ของแต่ละโมเดลเพื่อพัฒนาประสิทธิภาพของโมเดลให้ดีที่สุด

จากผลการทดลองพบว่าโมเดล MKML ให้ผลลัพธ์ที่มีประสิทธิภาพสูงที่สุด โดยประกอบไปด้วย Base Layer ทั้ง 6 อัลกอริทึม และนำผลลัพธ์ที่อยู่ในรูปแบบคะแนนความน่าจะเป็นในการเป็นโรคเบาหวานมาเป็นปัจจัยนำเข้าต่อใน Stack Layer ที่ใช้อัลกอริทึม Logistic Regression ในการสร้างโมเดลสุดท้าย โดยมีค่าความเที่ยงตรง (Accuracy) อยู่ที่ร้อยละ 74 แต่เนื่องจากคณะผู้จัดทำได้พบว่าโมเดลยังมีค่า Bias และ Variance ซึ่งอาจเป็นผลมาจากการจำนวนปัจจัยที่เกี่ยวข้องที่นำเข้าเพื่อสร้างโมเดลยังจำนวนที่น้อย ไม่ได้ทำลดความซ้ำซ้อนของข้อมูล และช่วงของค่าตัวแปรที่ปรับยังไม่กว้างมากพอ ส่งผลให้โมเดลมีโอกาสที่จะเกิด Overfitting และ Underfitting ได้ จึงจำเป็นต้องมีการพัฒนาโมเดลต่อไปในอนาคต

เอกสารอ้างอิง

[1] NCCPH. “Estimates of Diabetes and Its Burden in the United States.” **National Diabetes Statistics Report**. USA, 2017. pp.1-2.

[2]พญ.วรรณี นิธิยานันท์. **“คนไทยป่วย ‘เบาหวาน’ พุ่ง ป่วยแล้ว 5 ล้าน ทำให้เกิดโรคแทรกซ้อน พบบ่อย ‘ไตเรื้อรัง.’”** [Online]. เข้าถึงได้จาก : http://bit.ly/309LtG7. พ.ศ.2562.

[3]J. R. Quinlan. “Introduction of decision trees” **Machine Learning**. vol. 1, no. 1, March 1986. pp.81–106

[4]Peter Flach. **Machine Learning: The Art and Science of Algorithms that Make Sense of Data.** New York : Cambridge University Press. 2012.

[5]J. D. Kelleher, B. Mac Namee, and A. D’Arcy. **Fundamentals of machine learning for predictive data analytics : algorithms, worked examples, and case studies.** London : The MIT Press. 2015.

[6]R. Maclin and D. Opitz. “Popular Ensemble Methods: An Empirical Study” **J. Artificial Intelligence Reseach.** vol. 11, June 2011. pp.169–198.

[7] Somkiat. **“[Python] สรุป library เกี่ยวกับ Data Analysis สำหรับผู้เริ่มต้นไว้นิดหน่อย.”** [Online]. เข้าถึงได้จาก : http://bit.ly/2WtjjDX. พ.ศ.2562.

[8] CMU. Computer Science. **“NumPy and Matplotlib.”** [Silde]. เชียงใหม่ : ม.ป.ป.

[9]Kissada P. **“[ฝึกงาน] แยกประเภทรูปภาพด้วย Deep Learning ที่ Wongnai โจทย์ใหญ่ที่ไม่ธรรมดา.”** [Online] เข้าถึงได้จาก : http://bit.ly/2zFvy7i. พ.ศ.2562.

[10]C. Zhou, and others. “Learning Deep Representations from Heterogeneous Patient Data for Predictive Diagnosis” **Proceedings of the 8th ACM-BCB ’17**. 2017, pp.115–123.

[11]C. L. Cole and B. R. King. “Using Machine Learning to Predict the Health of HIV-Infected Patients” **BCB’13.** 2007, pp.684–685.

[12]R. P. Lippmann, L. Kukolich, and D. Shahian. “Predicting the risk of complications in coronary artery bypass operations using neural networks” **Proceedings of the 7th International Conference on Neural Information Processing Systems.** 1994, pp.1055–1062.

[13] H. Wu and M. D. Wang. “Infer Cause of Death for Population Health Using Convolutional Neural Network” **ACM-BCB ’17**. 2017. pp. 526–535.

[14]ชื่นชนก เชื้อพันธุ์. **“ประสบการณ์คุณแม่เกือบตายเพราะแผลผ่าคลอดติดเชื้อในกระแสเลือด.”** [Online]. เข้าถึงได้จาก : http://bit.ly/2JsyXvK. พ.ศ.2562.