

การรวมหลายโมเดลสำหรับทำนายความเสี่ยงโรคเบาหวาน
โดยใช้การเรียนรู้แบบมัลติเคอร์เนล
Combining Multiple Models for Predicting the Risk of Diabetes
Using Multi-Kernel Learning

ชนัดดาภรณ์ อ่อนแสง(Chanatdaporn Onsang)¹ ธนมณต์ ตันศิริเจริญกุล(Dhanamon Tansirisernkul)²

และธีรพงศ์ ลีลานูปา(Teerapong Leelanupab)³

¹ภาควิชาเทคโนโลยีสารสนเทศ ²ภาควิชาเทคโนโลยีสารสนเทศ ³ภาควิชาเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

¹58070025@it.kmitl.ac.th, ²58070053@it.kmitl.ac.th, ³teerapong@it.kmitl.ac.th

บทคัดย่อ

ในปัจจุบันมีผู้ป่วยหลายรายประสบกับโรคไม่ติดต่อเรื้อรัง (NCDs) เช่น โรคเบาหวาน, โรคหลอดเลือดหัวใจ และ โรคมะเร็ง ซึ่งการวินิจฉัยความผิดปกติของร่างกายนั้น จำเป็นต้องใช้ความพยายามอย่างมากในการอ่านและทำความเข้าใจผลการทดสอบจากห้องปฏิบัติการ โรคเบาหวานถือเป็นหนึ่งในโรคประเภท NCDs หลัก ๆ ซึ่งได้รับการรายงานมาเป็นเวลานานว่าเป็นปัญหาสุขภาพอันดับต้น ๆ ในหลายประเทศ การตระหนักถึงสัญญาณแรกเริ่มของโรคเบาหวานมีความสำคัญเป็นอย่างมากที่จะลดความเสี่ยงในการเกิดโรคแทรกซ้อนต่าง ๆ ที่จะตามมาในภายหลัง เช่น โรคไต, โรคหลอดเลือดสมอง, โรคความดันโลหิตสูง เป็นต้น ในงานวิจัยนี้ผู้วิจัยนำเสนอการศึกษาของการใช้เทคนิคการเรียนรู้แบบมัลติเคอร์เนลซึ่งรวมโมเดลการทำนายหลายแบบมาใช้ในการประเมินความเสี่ยงของโรคเบาหวาน ผลการทดลองพบว่าโมเดลการเรียนรู้แบบมัลติเคอร์เนลช่วยเพิ่มประสิทธิภาพค่าความเที่ยงตรงของโมเดลการทำนายแบบเดี่ยวทั้ง 5 รูปแบบ อยู่ที่ 1% - 7% และโมเดลการเรียนรู้แบบมัลติเคอร์เนลยังสามารถนำไปใช้เพื่อช่วยเหลือแพทย์ในการวินิจฉัยผู้ป่วยได้แม่นยำยิ่งขึ้นอีกด้วย

คำสำคัญ: การรวบรวมและจัดเตรียมข้อมูล, การประมวลผลข้อมูลขั้นต้น, การจัดหมวดหมู่และการจัดกลุ่ม, การแสดงแบบจำลองและการทำนาย

Abstract

At present, many patients suffer from noncommunicable diseases (NCDs), such as diabetes, cardiovascular disease and cancer, which for diagnosing body disorders require a large amount of effort to read and understand the test results from laboratories. One of the major NCDs is the diabetes that has been reported to be the top health problem in many countries. Recognizing the early signs of diabetes is of importance to reduce the risk of patients in exposing to other following complications, such as kidney failure, stroke, hypertension, etc. In this paper, we present a study of using the technique of multi-kernel learning in combining multiple predictive models in estimating the risk of diabetes. The experimental results show that our multi-kernel learning improves the accuracy of five individual predictive models from 1% - 7%. Our multi-kernel learning model can be used to assist physicians in diagnosing the risk of diabetes in patients more precisely.

Keyword: Data Acquisition and Preparation, Data Pre-processing, Classification and Clustering, Modeling and Prediction

1. บทนำ

จากรายงานสถิติโรคเบาหวานแห่งชาติปี 2558 มีผู้ป่วยโรคเบาหวาน 30.3 ล้านคนในสหรัฐอเมริกา หรือมีส่วนเท่ากับ 1 ใน 10 คนของประชากร ซึ่งได้รับการวินิจฉัยเป็นโรคเบาหวานแล้ว 23.1 ล้านคน และยังไม่ได้รับการวินิจฉัยทั้งหมด 7.2 ล้านคน จาก 30.3 ล้านคน มีผู้ป่วยจำนวน 30.2 ล้านคน มีเปอร์เซ็นต์การเพิ่มขึ้นของการเป็นโรคเบาหวานแปรผันกับอายุสูงถึง 25.2% ในอายุ 65 ปีขึ้นไป [1] สำหรับประเทศไทยจากรายงานของสำนักนโยบายและยุทธศาสตร์ สำนักงานปลัดกระทรวงสาธารณสุข พบว่าอัตราตายด้วยโรคเบาหวานต่อประชากรแสนคน ในภาพรวมของประเทศในปี 2556-2558 เท่ากับ 14.93, 17.53 และ 17.83 ตามลำดับ และจากการสำรวจสุขภาพประชาชนไทยอายุ 15 ปี ขึ้นไปโดยการตรวจร่างกายครั้งที่ 5 เมื่อปี 2557 พบว่า ความชุกของโรคเบาหวานเพิ่มสูงขึ้นเป็นร้อยละ 8.9 คิดเป็นจำนวนมากถึง 4.8 ล้านคน เทียบกับปี 2552 [2] จากสถิติพบว่าโรคเบาหวานเป็นปัญหาใหญ่ที่ต้องเผชิญอยู่ปัจจุบันนี้และส่งผลกระทบต่อผู้ป่วยทั้งด้านร่างกายและจิตใจ รวมถึงสังคมและวงการแพทย์อีกด้วย คณะผู้วิจัยได้ตั้งวัตถุประสงค์ของการทำงานวิจัยนี้ว่า

- 1) เพื่อค้นหา และบรรยายลักษณะอื่น ๆ ที่อาจเป็นปัจจัยส่งผลต่อการเป็นโรคเบาหวานประเภทที่ 2
- 2) เพื่อสร้างและพัฒนาโมเดลการทำนายให้มีประสิทธิภาพมากกว่าการใช้โมเดลการเรียนรู้แบบพื้นฐานที่นิยมใช้ทั่วไป
- 3) เพื่อพัฒนาโมเดลการเรียนรู้ที่สามารถช่วยในการประเมินความเสี่ยงการเป็นโรคเบาหวานประเภทที่ 2

จากการศึกษางานวิจัยอื่นก่อนหน้า พบว่าหลักการเรียนรู้ของเครื่อง (Machine Learning) ถูกนำมาใช้ในพัฒนาโมเดลสำหรับทำนายชุดข้อมูลโรคเบาหวานประเภทที่ 2 โดยใช้อัลกอริทึมประเภทการจัดหมวดหมู่ (Classification) [3] ซึ่งส่วนใหญ่นิยมใช้เทคนิคการเรียนรู้แบบโมเดลหรือแบบจำลองแบบเดี่ยว (Individual Standard Model) เช่น Logistic Regression หรือ Decision Tree [4] บางงานวิจัยใช้เทคนิคที่มีความซับซ้อนมากขึ้นโดยใช้เทคนิคการเรียนรู้แบบโมเดลหรือแบบจำลองแบบกลุ่ม (Ensemble) โดยเทคนิคแบบพื้นฐานของโมเดลแบบกลุ่มนั้นมี 2 ประเภท ได้แก่ Bagging และ Boosting [5] ตัวอย่างเช่น Random Forest และ XGBoost อย่างไรก็ตาม

โมเดลแบบกลุ่มเหล่านี้มีข้อจำกัด ได้แก่ ประเภทของโมเดลที่ใช้ต้องเป็นประเภทเดียวกัน และส่วนใหญ่อยู่ในไลบรารี (Library) สำเร็จรูปที่นิยมใช้ในงานเรียนรู้ของเครื่อง

งานวิจัยนี้นำเสนอการใช้การเรียนรู้แบบมัลติเคอร์เนลแบบหลายชั้น (Multiple Kernels on Multiple Layers -MKML) ซึ่งเป็นโมเดลการเรียนรู้แบบแบบกลุ่มประเภทหนึ่ง แต่มีหลายชั้นหรือเรียกว่า Stacking Ensemble Model ซึ่งยังไม่ถูกนำมาวิจัยอย่างแพร่หลาย อีกทั้งยังสามารถเพิ่มประสิทธิภาพความเที่ยงตรงของการทำนายโมเดลเหนือโมเดลแบบเดี่ยวและแบบกลุ่มแบบดั้งเดิม ซึ่งงานวิจัยนี้ทำการทดลองสร้างโมเดล MKML โดยทำการปรับแต่งและคัดเลือกประเภทโมเดลแบบเดี่ยวที่อยู่ภายในแต่ละชั้นของ MKML ที่ดีที่สุด และนำมาเปรียบเทียบประสิทธิภาพการทำนายของโมเดล MKML กับโมเดลการทำนายแบบเดี่ยวและแบบกลุ่ม ชนิดต่าง ๆ โดยใช้ชุดข้อมูลการตรวจโรคเบาหวานประเภทที่ 2

2. วรรณกรรมที่เกี่ยวข้อง

การเรียนรู้ด้วยเครื่องจักร (Machine Learning) คือ เทคนิคหนึ่งที่สำคัญในด้านวิทยาการข้อมูล (Data Science) ซึ่งทำให้คอมพิวเตอร์สามารถเรียนรู้ได้ด้วยตนเอง โดยรูปแบบของการเรียนรู้ [6] ที่นำมาใช้คือ Supervised Learning เป็นการเรียนรู้ที่เน้นสอนคอมพิวเตอร์โดยการศึกษาจากข้อมูลตัวอย่าง (Label) สามารถแบ่งประเภทอัลกอริทึม (Algorithm) ออกเป็น 2 กลุ่มได้แก่ การจัดหมวดหมู่ข้อมูล (Classification) และการถดถอย (Regression) และเป้าหมายของงานวิจัยคือสามารถทำนายความเสี่ยงว่าผู้ป่วยเป็นหรือไม่เป็นโรคเบาหวาน จึงใช้อัลกอริทึมประเภทการจัดหมวดหมู่ข้อมูล ซึ่งสามารถแบ่งออกได้เป็น 2 กลุ่มย่อย ตามวิธีการสร้างโมเดลจากอัลกอริทึม ได้แก่

2.1 เรียนรู้แบบเดี่ยว (Individual standard model)

หมายถึง แบบจำลองการเรียนรู้จากเพียงอัลกอริทึมเดียว โดยไม่ใช้ควบคู่กับตัวอื่น ซึ่งในงานวิจัยนี้ใช้ทั้งหมด 4 อัลกอริทึม ได้แก่

2.1.1 Decision tree [7]

เป็นแบบจำลองทางคณิตศาสตร์ที่เรียนรู้ข้อมูลเพื่อหาทางเลือกที่ดีที่สุดโดยอยู่ในรูปแบบเงื่อนไข และผลลัพธ์ ซึ่งเงื่อนไขดังกล่าวจะแบ่งข้อมูลตามคุณลักษณะ (Feature) เพื่อนำไปสร้างโมเดลในการทำนาย

¹ สามารถดาวน์โหลดชุดข้อมูลได้ที่ <https://www.kaggle.com/uciml/pima-indians-diabetes-database>

2.1.2 K-Nearest Neighbors [8]

เป็นวิธีการค้นหาเพื่อนบ้านที่ใกล้ที่สุด หรือข้อมูลที่มีความคล้ายคลึงกันมากที่สุดเป็นจำนวน k ตัว โดยใช้การคำนวณระยะห่างระหว่างข้อมูลกลุ่มตัวอย่างกับข้อมูลที่กำลังสนใจ ซึ่งการคำนวณวิธีต่าง ๆ ที่ขึ้นอยู่กับประเภทของคุณลักษณะข้อมูลและการนำไปใช้

2.1.3 Logistic Regression [9]

เป็นการประมาณการประเภทของข้อมูล โดยใช้ Logistic function สร้าง S-shaped curve ลากตัดผ่านกลุ่มข้อมูลตัวอย่างเพื่อทำนายว่าเป็นข้อมูลประเภทไหน ซึ่งเหมาะกับการทำนายข้อมูลประเภทจัดกลุ่ม

2.1.4 Support Vector Machine [10]

เป็นการจัดหมวดหมู่เชิงเรขาคณิตซึ่งคำนวณหาไฮเปอร์เพลน (Hyperplane) หรือระนาบที่มีขอบเขตน้อย เพื่อใช้ในการแบ่งหมวดหมู่ของข้อมูลในเวกเตอร์สเปซหลายมิติ โดยไฮเปอร์เพลนที่ดีที่สุดคือ ไฮเปอร์เพลนที่สามารถแบ่งข้อมูลออกจากกันได้มากที่สุดเมื่อเทียบกับระยะห่างจากไฮเปอร์เพลน

2.2 เรียนรู้แบบกลุ่ม (Ensemble model)

เป็นเทคนิคสร้างโมเดลจากหลายอัลกอริทึมมารวมผลลัพธ์การทำนายของการสร้างโมเดลแบบเดี่ยว ที่ได้ผลลัพธ์แตกต่างกันออกไปมาใช้ร่วมกันเพื่อเพิ่มประสิทธิภาพของโมเดล โดยที่ Ensembles มีหลักการทำงานพื้นฐาน แบ่งออกเป็นได้ 3 รูปแบบ ได้แก่ Bagging, Boosting และ Stacking [10] โดยในงานวิจัยนี้ได้เลือก Random Forest เป็นตัวอย่างของโมเดลการเรียนรู้แบบกลุ่มเพื่อมาเปรียบเทียบกับ Multiple Kernels on Multiple Layers (MKML)

2.2.1 Random Forest [5]

เป็นโมเดลการเรียนรู้แบบกลุ่มที่ใช้เทคนิค Bagging ที่สุ่มสร้างข้อมูลทดสอบ (Training sets) ที่แตกต่างกัน โดยการสุ่มเลือกฟีเจอร์เพื่อนำมาใช้กับโมเดล ซึ่งผลลัพธ์คือการสร้างโมเดล Decision Tree ขึ้นมาหลายโมเดล และนำผลการทำนายมารวมกันโดยวิธีการโหวต

2.2.2 Multiple Kernels on Multiple Layers (MKML) [11]

เป็นโมเดลหลักที่นำเสนอในงานวิจัยนี้ ซึ่งใช้เทคนิครูปแบบ Stacking ที่เป็นการใช้การเรียนรู้แบบกลุ่ม โดยรวมอัลกอริทึมการทำนายที่แตกต่างกันมากกว่า 1 ประเภทมารวมไว้ในที่เดียวและมีการขั้นตอนการทำนายมากกว่า 1 ชั้น และมีการใช้ K-fold

Cross validation เพื่อสร้างการรวมการถ่วงน้ำหนักที่ดีที่สุดของการทำนาย อีกทั้งยังสามารถหลีกเลี่ยงการเกิดโมเดลที่ใช้ในการทำนายได้ดีกับข้อมูลฝึกมากเกินไป และสมมติฐานที่ไม่เหมาะสม

นอกจากการศึกษางานวิจัยอื่นที่เกี่ยวข้องในเรื่องการสร้างโมเดล คณะผู้วิจัยได้ให้ความสำคัญกับขั้นตอนการทำความสะอาดข้อมูลเพื่อเพิ่มประสิทธิภาพการทำนายของโมเดล และจากการศึกษางานวิจัยอื่นพบว่าใช้วิธีการทำความสะอาดข้อมูลโดยใช้ค่าเฉลี่ยของคอลัมน์นั้น ๆ เช่น งานวิจัยของ S. Priya และ R.R.Rajalaxmi [12] ได้ทำความสะอาดข้อมูลเฉพาะค่า ค่าสูญหายโดยใช้วิธีการแทนค่าเฉลี่ย และใช้เทคนิคการจำแนกข้อมูลด้วย Neural Network ซึ่งได้ผลประสิทธิภาพค่าความเที่ยงตรงที่ดีที่สุดอยู่ที่ 0.97 จากงานวิจัยดังกล่าวผู้วิจัยจึงได้นำมาศึกษาหลักการทำความสะอาดข้อมูล ควบคู่ไปกับประสิทธิภาพการทำนายของโมเดล และตัดสินใจที่จะหาวิธีที่แตกต่างออกไปมาช่วยเพิ่มประสิทธิภาพของโมเดล

3. วิธีการดำเนินงานวิจัย

3.1 การจัดหาและทำความเข้าใจข้อมูล

ผู้วิจัยได้ศึกษาการวินิจฉัยโรคเบาหวานเบื้องต้นโดยสมาคมโรคเบาหวานแห่งสหรัฐอเมริกา (ADA) กล่าวว่าปัจจัยสำคัญที่ใช้ในการวินิจฉัยคือค่าระดับน้ำตาลในพลาสมาทุกสัปดาห์อดอาหาร [13] จากนั้นผู้วิจัยจึงได้ตั้งข้อสังเกตว่านอกจากค่ากลูโคส ยังมีปัจจัยอื่นใดที่อาจส่งผลให้เป็นโรคเบาหวานได้อีก จึงเริ่มทำการจัดหาชุดข้อมูลเบาหวานประเภทที่ 2 Pima Indian Diabetes ซึ่งเป็นข้อมูลของผู้หญิงชาวอเมริกันพื้นเมือง หรือที่เรียกกลุ่มนี้ว่าชาว Pima Indian มีอายุ 20 ปีขึ้นไป¹ ที่ได้เข้ารับการตรวจวินิจฉัยโรคเบาหวานประเภทที่ 2 ซึ่งข้อมูลได้เผยแพร่ที่เว็บไซต์ของสถาบันโรคเบาหวานและโรคทางเดินอาหารและโรคไตแห่งชาติ (National Institute of Diabetes and Digestive and Kidney Diseases - NIDDK) ของประเทศสหรัฐอเมริกา [14] โดยประกอบด้วยข้อมูล 2 กลุ่ม คือ กลุ่มที่ไม่เป็นเบาหวาน มี 500 คน และกลุ่มที่เป็นโรคเบาหวาน 276 คน รวมทั้งหมด 768 คน และมีฟีเจอร์ทั้งหมด 8 ตัวแปร โดยมีรายละเอียดดังตารางที่ 1

¹ สามารถดาวน์โหลดชุดข้อมูลได้ที่ <https://www.kaggle.com/uciml/pima-indians-diabetes-database>

ตารางที่ 1 : พจนานุกรมข้อมูล Pima Indian Diabetes

ชื่อ	คำอธิบาย
Number of times pregnant	จำนวนครั้งที่ตั้งครรภ์
Plasma glucose	ความเข้มข้นของกลูโคสในช่องปาก
Diastolic blood pressure	ความดันโลหิต
Triceps skin fold thickness	ความหนาของไขมันใต้ผิวหนัง
2-Hour serum insulin (Mu U/ml)	ความเข้มข้นของอินซูลินในร่างกายหลังฉีดเซรัมเข้าไป 2 ชม.
BMI	ดัชนีมวลกาย
Diabetes pedigree function	ความเสี่ยงของโรคเบาหวานที่มาจากทางพันธุกรรม
Age	อายุ
Has Diabetes	เป็นโรคเบาหวานประเภทที่ 2 หรือไม่

3.2 การเตรียมข้อมูล

เนื่องจากข้อมูลที่ได้มาอาจมีปัญหาเรื่องข้อมูลผิดพลาด (Noisy Data) เช่น มีค่าผิดปกติ (Outlier), ค่าสูญหาย (Missing Values) และข้อมูลไม่สอดคล้องกัน (Inconsistent Data) จึงจำเป็นต้องแก้ไขหรือปรับข้อมูลเหล่านี้ให้กลับมาเป็นสภาพปกติ คณะผู้วิจัยจึงเริ่มจากการตรวจจับข้อมูลที่มีความผิดพลาดจากการเก็บข้อมูลโดยใช้ Interquartile range (IQR) ส่วนของค่าสูญหาย เริ่มจากการหาค่า NaN หรือข้อมูลที่ไม่น่าจะเป็นค่า 0 ด้วย Boxplot จากนั้นจึงทำการแก้ไขข้อมูลที่ผิดพลาด โดยผู้วิจัยได้ศึกษาจากผู้วิจัยรายอื่นที่เคยทำสร้างโมเดลกับข้อมูลชุดเดียวกัน ซึ่งได้ทำการแก้ไขด้วยการแทนค่าเฉลี่ยของข้อมูลแต่ละฟีเจอร์ แต่จากการศึกษาเพิ่มเติมได้พบว่ายังมีอีกหลากหลายวิธีที่สามารถแทนค่าผิดพลาดดังกล่าว เช่น การลบข้อมูลแถวนั้นทิ้งไปเลย หากคอลัมน์นั้น ๆ มีจำนวนค่าผิดปกติน้อย จากการศึกษาดังกล่าวคณะผู้วิจัยจึงสนใจการทดลองแทนค่าผิดปกติหรือค่าสูญหายเหล่านั้นด้วยค่าเฉลี่ย

3.3 การคัดเลือกปัจจัยสำคัญในการสร้างโมเดล

ใช้วิธีการคัดเลือกฟีเจอร์โดยการคำนวณค่าสหสัมพันธ์ของเพียร์สัน (Pearson Correlation) เพราะเหมาะกับการหาค่าสหสัมพันธ์ที่เป็นรูปแบบตัวเลขที่เป็นลักษณะข้อมูลเชิงปริมาณแบบต่อเนื่อง (Continues Data) ข้อมูลระดับช่วงชั้น (Interval Scale) และข้อมูลระดับอัตราส่วน (Ratio Scale) ซึ่งจะได้ค่าสหสัมพันธ์ดังตารางที่ 2

ตารางที่ 2 : ค่าสหสัมพันธ์เพียร์สันของแต่ละฟีเจอร์เทียบกับเลเบล

ชื่อ	ค่าสหสัมพันธ์
Number of times pregnant	0.22
Plasma glucose	0.49
Diastolic blood pressure (mm Hg)	0.17
Triceps skin fold thickness	0.07
2-Hour serum insulin (Mu U/ml)	0.08
BMI	0.3
Diabetes pedigree function	0.17
Age	0.24

จากตารางที่ 2 พบว่าค่าสหสัมพันธ์ของฟีเจอร์กับเลเบลมีความสัมพันธ์ไปในทิศทางเดียวกัน คณะผู้วิจัยจึงได้เลือก 4 ปัจจัยสำคัญ ได้แก่ Plasma glucose, Age, BMI และ Number of times pregnant มาเป็นฟีเจอร์ในการทำนาย เพราะมีค่าใกล้เคียง 1 มากที่สุดเมื่อเทียบกับฟีเจอร์อื่น ๆ

3.4 การออกแบบการทดลอง

เริ่มจากการแบ่งชุดข้อมูลด้วยวิธี K-fold cross validation (CV) ออกเป็นจำนวน k ส่วน จำนวนเท่า ๆ กัน ซึ่งในแต่ละส่วนจะถูกแบ่งให้มีข้อมูลฝึก (Training Set) และข้อมูลทดสอบ (Test Set) และในการทำงานข้อมูลทั้ง k ส่วนจะมีสลับหน้าที่เป็นข้อมูลฝึก และข้อมูลทดสอบ และทำงานจนครบ k รอบ ซึ่งจะสร้างทั้งหมด k โมเดลที่แตกต่างกัน โดยการทดลองนี้กำหนดค่า k = 10 ดังนั้นในแต่ละรอบข้อมูลฝึกจะมีขนาด 90% ของข้อมูลทั้งหมด และข้อมูลทดสอบมีขนาด 10%

เพื่อเปรียบเทียบประสิทธิภาพของ MKML การทดลองนี้ได้เลือกโมเดลประเภทต่าง ๆ ที่เป็นทั้งแบบเดี่ยวและแบบกลุ่ม ซึ่งประกอบไปด้วย Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbor (KNN) และ Random Forest (RF) จากนั้นได้ทำการปรับพารามิเตอร์ของแต่ละโมเดลโดยใช้วิธี Grid Search โดยกำหนดให้ CV = 10 เพื่อให้ได้โมเดลมีค่าความเที่ยงตรง (Accuracy) สูงสุด

3.5 การพัฒนาโมเดล MKML

MKML แบ่งการฝึกข้อมูลออกเป็น 2 ชั้น ในขั้นแรกมีการกำหนดทั้งหมด 6 อัลกอริทึม ได้แก่ Decision Tree, Logistic Regression, K-Nearest Neighbors, Logistic RegressionCV,

¹ สามารถดาวน์โหลดชุดข้อมูลได้ที่ <https://www.kaggle.com/uciml/pima-indians-diabetes-database>

Random Forest และ Support Vector Machine และในขั้นที่สอง จะกำหนดได้เพียง 1 อัลกอริทึมเท่านั้นจึงได้ทำการทดลอง 4 อัลกอริทึมเพื่อทำการคัดเลือกโมเดลที่ดีที่สุด ได้แก่ Decision Tree, Support Vector Machine, Logistic Regression และ Random Forest โดยที่ข้อมูลนำเข้าของขั้นแรกคือ Descriptive feature จากนั้นนำผลลัพธ์ของขั้นแรกมาเป็นข้อมูลนำเข้าของ ขั้นที่สอง และผลลัพธ์ของขั้นที่สองคือผลลัพธ์การทำนายของ MKML ซึ่งโมเดลจะได้รับการปรับแต่งคัดเลือกพารามิเตอร์ที่ทำให้ MKML มีประสิทธิภาพสูงสุด ประกอบไปด้วย ขั้นที่หนึ่ง คือทั้ง 6 อัลกอริทึมเหมือนเดิม และขั้นที่สองคือ Logistic Regression

3.6 การวัดประสิทธิภาพของโมเดล

เป็นขั้นตอนทดสอบประสิทธิภาพโมเดลจากอัลกอริทึมที่ใช้ในการทำนาย โดยตัวอย่างของตัวชี้วัดในด้วยการเรียนรู้ด้วยเครื่องจักร มีดังต่อไปนี้ [15]

3.6.1 ค่าความเที่ยงตรง (Accuracy)

คำนวณค่าที่ทำนายถูกต้องเพื่อหาความเที่ยงตรงของโมเดล โดยคิดจากสัดส่วนของจำนวนการทำนายที่ถูกต้องเทียบกับจำนวนการทำนายทั้งหมด ซึ่งในการเรียนรู้ของเครื่องให้ความสำคัญกับค่าความเที่ยงตรงของข้อมูลทดสอบ (Test Accuracy) คือความเที่ยงตรงที่ทำการทดลองกับข้อมูลทดสอบ (Test Set) มากที่สุด อย่างไรก็ตามในงานวิจัยนี้ยังรายงานผลค่าความเที่ยงตรงของข้อมูลฝึก คือความเที่ยงตรงที่วัดเมื่อทำการทดลองกับข้อมูลฝึกทั้งนี้เพื่อนำมาใช้ในการคำนวณค่าเอนเอียงและค่าแปรปรวน

3.6.2 ความแม่นยำ (Precision)

คำนวณหาการทำนายที่ถูกต้องเพื่อหาความแม่นยำของโมเดล โดยคิดจากสัดส่วนของจำนวนการทำนายว่าเป็นโรคถูกต้อง (True positive) เทียบกับผลรวมของจำนวนการทำนายว่าเป็นโรคที่ถูกต้อง และการทำนายว่าไม่เป็นโรคผิดเป็นโรค (False positive)

3.6.3 ค่าเอนเอียง (Bias)

เป็นการวัดค่าเฉลี่ยของค่าที่ทำนายออกมาได้ มีความใกล้เคียงกับข้อมูลจริงมากน้อยเพียงใด ยังมีค่าสูงหมายความว่าค่าทำนายไม่ใกล้เคียงกับค่าความจริง (Underfitting)

3.6.4 ค่าแปรปรวน (Variance)

เป็นการวัดค่าความผิดพลาด (Error) ของความสามารถในการทำนายโมเดล เมื่อมีการเปลี่ยนแปลง ข้อมูลทดสอบ หากมีแปรปรวนต่ำจะไม่เกิดปัญหาการเกิดอาการโมเดลที่ใช้ในการทำนายได้ดีกับข้อมูลฝึกมากเกินไป (Overfitting)

4. ผลการดำเนินงานวิจัย

4.1 ผลและวิเคราะห์การเปรียบเทียบประสิทธิภาพโมเดลการทำนายโรคเบาหวานแบบเดี่ยว และแบบกลุ่ม ในชุดข้อมูลฝึก และทดสอบ

ตารางที่ 3 : แสดงผลการเปรียบเทียบประสิทธิภาพความเที่ยงตรง ค่า

Algorithm	Individual Standard Model					Ensemble Model
	DT	KNN	SVM	LR	RF	MKML
Training Accuracy	0.99	0.81	0.77	0.77	1.00	0.85
Test Accuracy	0.70	0.75	0.77	0.77	0.76	0.78
Bias	0.01	0.19	0.23	0.23	0.00	0.15
Variance	0.30	0.06	0.00	0.00	0.24	0.07
Precision	0.62	0.69	0.73	0.73	0.68	0.75

เอนเอียง ค่าแปรปรวน และค่าความแม่นยำของโมเดลทำนายโรคเบาหวาน ในชุดข้อมูลฝึก และทดสอบ

จากตารางที่ 3 การเปรียบเทียบประสิทธิภาพของโมเดลการทำนายโรคเบาหวานที่อยู่ในรูปแบบการเรียนรู้แบบเดี่ยวจากเทคนิคการจำแนกประเภทข้อมูลทั้ง 5 อัลกอริทึม และแบบกลุ่มด้วยค่าความเที่ยงตรง, ค่าเอนเอียง, ค่าแปรปรวน และค่าความแม่นยำ มีค่าความเที่ยงตรงของข้อมูลฝึก (Accuracy of Training Set – Training Accuracy) สูงสุดอยู่ที่ 1.00 ของ RF และค่าความเที่ยงตรงของข้อมูลทดสอบ (Accuracy of Test Set – Test Accuracy) สูงสุดอยู่ที่ 0.78 เป็นของ MKML ส่วนค่าเอนเอียงและค่าแปรปรวน มีค่าต่ำสุดอยู่ที่ 0.00 เป็นของ Random Forest และ Logistic Regression

นอกจากนี้ยังมีการตรวจค่าความแม่นยำ (Precision) มีค่าสูงสุดอยู่ที่ 0.75 ซึ่งเป็นของ MKML จากการเปรียบเทียบ

¹ สามารถดาวน์โหลดชุดข้อมูลได้ที่ <https://www.kaggle.com/uciml/pima-indians-diabetes-database>

ประสิทธิภาพทั้ง 4 รูปแบบ พบว่าเทคนิค MKML นั้นให้ประสิทธิภาพที่ดีกว่าการเรียนรู้แบบเดี่ยวทั้ง 5 รูปแบบ

5. สรุป

ในปัจจุบันโรคเบาหวานเป็นประเด็นปัญหาอันดับต้นที่สังคมกำลังเผชิญอยู่จากสถิติการเกิดโรคเบาหวานของประชาชนที่มีแนวโน้มเพิ่มสูงขึ้นทุกปี ซึ่งงานวิจัยนี้ได้มีการทดลองใช้เทคนิค Multiple Kernels on Multiple Layers (Stacking Ensemble Model) มาช่วยในการสร้างและพัฒนาโมเดลในการทำนายการเกิดโรคเบาหวาน เพื่อเป็นการป้องกันและเฝ้าระวังการเกิดโรคเบาหวานตั้งแต่เนิ่น ๆ ในขั้นตอนของการพัฒนาโมเดล ทางผู้วิจัยได้เปรียบเทียบประสิทธิภาพโมเดลระหว่างแบบเดี่ยวกับแบบกลุ่ม ซึ่งใช้เทคนิค MKML

จากผลการทดลอง พบว่าการใช้เทคนิค MKML ได้ค่าความเที่ยงตรงของ Test Accuracy ในการทำนายสูงกว่าแบบเดี่ยวทั้ง 5 รูปแบบ โดยเทียบกับค่าความเที่ยงตรงของข้อมูลทดสอบของโมเดลในขั้นตอน Individual Standard Model ที่มีค่าความเที่ยงตรงน้อยที่สุด คือ Decision Tree มีค่าเพิ่มขึ้น 8.41% และเปรียบเทียบกับอัลกอริทึมที่มีค่าความเที่ยงตรงมากที่สุด คือ Logistic Regression มีค่าเพิ่มขึ้น 1.24% สิ่งที่ผู้วิจัยต้องการแก้ไขและพัฒนาต่อไปในอนาคต คือ เพิ่มประสิทธิภาพในการทำนายของเทคนิค MKML ให้มีประสิทธิภาพดีขึ้น นอกจากนี้ยังต้องการขยายขอบเขตในการทำนายการเกิดโรคต่าง ๆ เพิ่มเติม

เอกสารอ้างอิง

- [1] National Center for Chronic Disease Prevention and Health Promotion, "Estimates of Diabetes and Its Burden in the United States." National Diabetes Statistics Report, pp.1-2, 2017.
- [2] พญ.วรรณิ นิธิยานันท์, "คนไทยป่วย 'เบาหวาน' พุง ป่วยแล้ว 5 ล้าน ทำให้เกิดโรคแทรกซ้อน พบบ่อย 'ไตเรื้อรัง.'" [Online]. เข้าถึงได้จาก : <https://www.hfocus.org/content/2016/11/12992>. พ.ศ.2562.
- [3] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," *Procedia Comput. Sci.*, vol. 132, pp. 1578–1585, 2018.
- [4] D. Choubey, S. Paul, S. Kumar, and S. Kumar, "Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection," in *Communication and Computing Systems*, pp. 451–455, 2016.
- [5] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques," *Front. Genet.*, vol. 9, 2018.
- [6] T. M. Mitchell, *Machine Learning*. New York, USA, 1997.
- [7] J. R. Quinlan, "Introduction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.
- [8] Peter Flach, *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. New York: Cambridge University Press, 2012.
- [9] J. D. Kelleher, B. Mac Namee, and A. D'Arcy, *Fundamentals of machine learning for predictive data analytics : algorithms, worked examples, and case studies*. Cambridge, Massachusetts, London: The MIT Press, 2015.
- [10] R. Maclin and D. Opitz, "Popular Ensemble Methods: An Empirical Study," *J. Artif. Intell. Res.*, vol. 11, pp. 169–198, Jun. 2011.
- [11] A. I. Naimi and L. B. Balzer, "Stacked Generalization: An Introduction to Super Learning." in *European Journal of Epidemiology*, 2017.
- [12] S. R. R. R. Priya, "An Improved Data Mining Model to Predict the Occurrence of Type-2 Diabetes using Neural Network," *IJCA*, 2012.
- [13] ณัฐธยาน์ ประเสริฐอำไพสกุล, "ปัจจัยทำนายพฤติกรรมป้องกันการภาวะน้ำตาลในเลือดสูงของผู้สูงอายุ โรคเบาหวานที่ไม่สามารถควบคุมระดับน้ำตาลในเลือด," กรุงเทพฯ: มหาวิทยาลัยธรรมศาสตร์, 2550.
- [14] J. W. Smith, J. E. Everhart, W. C. Dicksont, W. C. Knowler, and R. S. Johannes, "Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus." p.261-265, 1998.
- [15] Somkiat, "[Python] สรุปลibraryเกี่ยวกับ Data Analysis สำหรับผู้เริ่มต้นไว้วันไหน," [Online]. เข้าถึงได้จาก: <http://bit.ly/2PWBBeg/>. 2019.

¹ สามารถดาวน์โหลดชุดข้อมูลได้ที่ <https://www.kaggle.com/uciml/pima-indians-diabetes-database>