

คุณากรณ์ พันธุ์เพียร : การทำนายโรคเบาหวานโดยใช้วิศวกรรมคุณลักษณะสำหรับขั้นตอนวิธีการจำแนกในการเรียนรู้ของเครื่อง (DIABETIC PREDICTION USING FEATURE ENGINEERING FOR CLASSIFICATION ALGORITHM IN MACHINE LEARNING) อาจารย์ที่ปรึกษา : ผู้ช่วยศาสตราจารย์ ดร.เจษฎา ตัณฑนุช, 57 หน้า.

คำสำคัญ : โรคเบาหวาน, วิศวกรรมคุณลักษณะ, การเรียนรู้ของเครื่อง

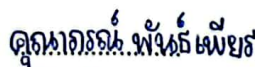
จุดมุ่งหมายของการศึกษานี้คือ เพื่อศึกษาปัจจัยเสี่ยงที่มีผลกระทบต่อการทำให้เกิดโรคเบาหวานโดยใช้วิธีวิศวกรรมคุณลักษณะ แล้วนำไปพัฒนาแบบจำลองเพื่อจำแนกประเภทของผู้ป่วยโรคเบาหวานและไม่เป็นโรคเบาหวาน ซึ่งการศึกษานี้ได้นำข้อมูลจากการตอบแบบสำรวจจำนวน 70,692 รายการ โดยได้ข้อมูลจากเว็บไซต์ Kaggle แล้วนำมาใช้วิธีวิศวกรรมคุณลักษณะเพื่อหาคุณลักษณะที่ดีที่สุดออกมา จากนั้นนำไปสร้างแบบจำลองด้วย 5 ขั้นตอนวิธี ได้แก่ เทคนิคป่าสุ่ม เทคนิคต้นไม้ตัดสินใจ เทคนิคเพื่อนบ้านใกล้ที่สุด เทคนิคเกรเดียนท์บูตทรี และซัพพอร์ตเวกเตอร์แมชชีน แล้วทำการเปรียบเทียบประสิทธิภาพและความแม่นยำของแบบจำลอง ผลการศึกษาพบว่าประสิทธิภาพการทำงานของบางแบบจำลองเมื่อเปรียบเทียบระหว่างชุดข้อมูลต้นฉบับกับชุดข้อมูลที่ผ่านการสกัดคุณลักษณะด้วยวิธีวิศวกรรมคุณลักษณะมีค่าลดลงแต่ยังคงมีค่าใกล้เคียงกันมาก ซึ่งได้แก่แบบจำลองที่สร้างโดยเทคนิคป่าสุ่มและเทคนิคต้นไม้ตัดสินใจ แต่แบบจำลองที่สร้างโดยเทคนิคเกรเดียนท์บูตทรี เทคนิคเพื่อนบ้านใกล้ที่สุด และซัพพอร์ตเวกเตอร์แมชชีน มีประสิทธิภาพที่มากขึ้นกว่าเดิมอย่างชัดเจน ทั้งนี้พบว่าการใช้วิธีวิศวกรรมคุณลักษณะร่วมกับการสร้างแบบจำลองด้วยเทคนิคป่าสุ่ม สามารถให้ประสิทธิภาพการทำงานโดยภาพรวมอยู่ในเกณฑ์ที่ดีกว่าแบบจำลองอื่น ๆ โดยคุณลักษณะที่สกัดออกมาได้นั้นเหลือเพียง 7 คุณลักษณะ จากทั้งหมด 21 คุณลักษณะ ได้แก่ 1) ภาวะความดันโลหิต 2) ภาวะคลอเลสเตอรอลสูง 3) ค่าดัชนีมวลกาย 4) ระดับสุขภาพโดยทั่วไป 5) เพศ 6) ระดับช่วงอายุ และ 7) ระดับเงินเดือน โดยมีผลการวัดประสิทธิภาพดังนี้ Accuracy = 73.35% Precision = 68.53% Specificity = 60.17% Sensitivity = 87.02% F1 score = 76.55% ROC AUC = 0.815 และ Kappa = 0.469 ดังนั้นในการศึกษานี้พบว่า การนำวิธีวิศวกรรมคุณลักษณะเข้ามาทำงานร่วมกับการสร้างแบบจำลองด้วยเทคนิคป่าสุ่ม ทำให้สามารถสร้างแบบจำลองเหมาะสมในการคัดกรองผู้ป่วยที่เป็นโรคเบาหวานได้ดีที่สุด

KUNAPORN PUNPAIN: DIABETIC PREDICTION USING FEATURE ENGINEERING FOR CLASSIFICATION ALGORITHM IN MACHINE LEARNING. THESIS ADVISOR : ASST. PROF. JESSADA TANTHANUCH, Ph.D., 57 PP.

Keyword : DIABETES, FEATURE ENGINEERING, MACHINE LEARNING

The aim of this study is to study the risk factors affecting the incidence of diabetes using feature engineering method and then to develop a model to classify the types of diabetic and non-diabetic patients. This study used data from 70,692 survey responses from Kaggle website. First the feature engineering method was applied to extract the best features and then created models by 5 algorithms, which were random forest technique (RFT), decision tree technique (DTT), nearest neighbor technique (NNT), gradient boot tree technique (BGT), and support vector Machine (SVM). All models were evaluated the performance and the accuracy. The results showed that the performance of some models compared between the original dataset and the feature extracted by feature engineering method was lower but still very close, i.e. the model generated by RFT and DTT. However, for the model created by GBT, NNT, and SVM, feature engineering method clearly helped in increasing the efficiency. It was found that the use of feature engineering method together with RFT for creating the model provided the better overall performance than other methods. For the model mentioned, only 7 of the 21 features extracted were available, namely: 1) high blood pressure 2) high cholesterol 3) BMI 4) general health 5) gender 6) age and 7) salary. The efficiency measurement results were as follows: Accuracy = 73.35% Precision = 68.53% Specificity = 60.17% Sensitivity = 87.02% F1 score = 76.55% ROC AUC = 0.815 and Kappa = 0.469. Hence, the result of the study presents that the use of feature engineering method together with RFT offers a suitable model for the best screening of patients with diabetes.

School of Biomedical Innovation Engineering
Academic Year 2021

Student's Signature 
Advisor's Signature