การศึกษาเทคนิคการทำเหมืองข้อมูล เพื่อพยากรณ์ความเสี่ยงการเกิดโรคเบาหวาน

A survey of Data Mining Techniques for Predictions of Diabetes

จตุพล จิตติยพล (Jatupol Jittiyapol)[1] ,ณัฐิสรา ชูลิขิต (Nutthisara Choolikhit)[2] พัชรินทร์ พลเยี่ยม (Patcharin Ponyiam)[3] และนลัทพร โอษฐิเวช(Naladtaporn Aottiwerch)[4]

[1]ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น

บทคัดย่อ

โรคเบาหวานเป็นอีก 1 โรคเรื้อรังที่มีผู้ป่วยเพิ่มมากขึ้น โรคเบาหวานเป็นโรคที่ติดอันดับ 1 ใน 5 โรคของเอเชีย และติดอันดับ 1 ใน 10 ของโรคที่คุกคามคนไทยมากที่สุด ในประเทศไทยพบว่ามีผู้เสียชีวิตจากโรคเบาหวานเฉลี่ยสูงถึงวันละ 27 ราย อีกทั้งยังเป็นโรคที่มีสถิติแนวโน้มของผู้ป่วยเพิ่มสูงขึ้นอย่างต่อเนื่องทุกๆ ปีอีกด้วย

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาเทคนิคการทำเหมืองข้อมูลที่เหมาะสมในการพยากรณ์ความเสี่ยงของการเกิดโรคเบาหวาน โดยทำการศึกษาเทคนิคต่างๆ จากงานวิจัยที่มีการนำเทคนิคเหมืองข้อมูลมาใช้ในการพยากรณ์หาสาเหตุและความเสี่ยงในการเกิดโรค เช่น เทคนิคต้นไม้ตัดสินใจ (Decision Tree) , โครงข่ายประสาทเทียมแบบแพร่กลับ (Backpropagation Neural Network: BPNN), การวิเคราะห์การถดถอยโลจิสติก (Logistic Regression Analysis), การแบ่งกลุ่ม (Cluster Analysis) และNaïve byes เป็นต้น

จากการศึกษาข้างต้นพบว่าค่าความถูกต้องในการพยากรณ์ของ เทคนิคต้นไม้ตัดสินใจ (Decision Tree) และโครงข่ายประสาทเทียมแบบแพร่กลับ (Backpropagation Neural Network: BPNN) มีความเหมาะสมในการนำไปใช้สร้างแบบจำลองพยากรณ์ความเสี่ยงของการเกิดโรคเบาหวาน

คำสำคัญ: Data Mining, Prediction, K-Mean, Logistic Regression, Decision Tree, Support Vector Machine (SVM), Naive Bayesian, Gaussian Processes (GP), SEM, Diabetes

Abstract

The purpose of this research was to compare the effectiveness of the data mining techniques for predictions of diabetes. This research study and comparison other research about predicting diabetes by risk factors. 7 techniques of models were K-Mean, Logistic Regression, Decision Tree, Support Vector Machine (SVM), Naive Bayesian, Gaussian Processes (GP) and .....(SEM). The results is Decision Tree have accuracy more than other techiques.

Keyword: Data Mining, Prediction, K-Mean, Logistic Regression, Decision Tree, Support Vector Machine (SVM), Naive Bayesian, Gaussian Processes (GP), SEM, Diabetes

1. บทนำ

ปัจจุบันทางด้านสาธารณสุขทั่วโลกต่างให้ความสำคัญกับการจัดการโรคเรื้อรังเพิ่มมากขึ้น โดยโรคเบาหวานเป็นอีก 1 โรคเรื้อรังที่มีผู้ป่วยเพิ่มมากขึ้น โรคเบาหวานเป็นโรคที่ติดอันดับ 1 ใน 5 โรคของเอเชีย และติดอันดับ 1 ใน 10 ของโรคที่คุกคามคนไทยมากที่สุด ในประเทศไทยพบว่ามีผู้เสียชีวิตจากโรคเบาหวานเฉลี่ยสูงถึงวันละ 27 ราย[7] อีกทั้งยังเป็นโรคที่มีสถิติแนวโน้มของผู้ป่วยเพิ่มสูงขึ้นอย่างต่อเนื่องทุกๆ ปีอีกด้วย

ดังนั้นจึงมีแนวคิดในการนำเทคโนโลยีการทำเหมืองข้อมูลมาใช้ในการพยากรณ์ โดยได้

ทำการศึกษาเทคนิคต่างๆ เพื่อพิจารณาเทคนิคที่เหมาะสมในการนำมาวิเคราะห์ข้อมูลการพยากรณ์ความเสี่ยงของการเกิดโรคเบาหวาน

2. ทฤษฎีและวรรณกรรมที่เกี่ยวข้อง

2.1 โรคเบาหวาน (Diabetes Mellitus)[1]

โรคเบาหวาน คือ ภาวะระดับน้ำตาลสูง เกิดจากตับอ่อนไม่สามารถผลิตอินซูลินออกมาได้เพียงพอกับความต้องการของ

ร่างกาย ทำให้ร่างกายไม่สามารถใช้อินซูลินได้อย่างมีประสิทธิภาพ เมื่อระดับน้ำตาลในเลือดสูงเป็นเวลานานๆ อาจจะทำให้เกิดโรคแทรกซ้อนต่างๆ ตามมา เช่น ตา ไต หัวใจ เท้า เป็นต้น

2.2 เหมืองข้อมูล (Data Mining)[3]

คือ กระบวนการที่กระทำกับข้อมูลจำนวนมากเพื่อค้นหารูปแบบและความสัมพันธ์ ที่ซ่อนอยู่ในชุดข้อมูลนั้นๆ มีขั้นตอนดังนี้

- ทำความเข้าใจปัญหา
- ทำความเข้าใจข้อมูล
- เตรียมข้อมูล
- สร้างแบบจำลอง
- ประเมินผล
- นำไปใช้งาน

2.3 ต้นไม้ตัดสินใจ (Decision Tree)[3]

เป็นการเรียนรู้โดยการจำแนก (Classification) ข้อมูลเป็นกลุ่ม (Class) ต่างๆ โดยใช้คุณสมบัติ (Attribute) ของข้อมูลใน

การจำแนกต้นไม้ตัดสินใจประกอบด้วยอัลกอริทึม เช่น ID3,C4.5 เป็นต้น โดยมีโครงสร้างเป็นต้นไม้แตกแขนงไปตามเงื่อนไขหรือข้อมูลที่ได้คาดคะเนไว้ว่าจะเกิดขึ้น

โครงสร้างของ Decision tree จะประกอบด้วย

1. ใบ (Leaf) เป็นส่วนของผลลัพธ์ที่ได้
2. กิ่ง (Branch) เป็นตัวเชื่อมระหว่างโหนด
3. ราก (Root) เป็นจุดเริ่มต้นของเหตุการณ์

2.4 โครงข่ายประสาทเทียมแบบแพร่กลับ (Backpropagation Neural Network: BPNN)[6]

เป็นโมเดลทางคณิตศาสตร์ สำหรับประมวลผลสารสนเทศด้วยการคำนวณแบบ Connectionist เพื่อจำลองการทำงานของเครือข่ายประสาท คล้ายกับการทำงานในสมองมนุษย์ โดยคำนวณหาความสัมพันธ์ระหว่างปัจจัยนำเข้า (Input) และผลลัพธ์ (Output) การเรียนรู้นี้จะเริ่มจากการสุ่มค่าน้ำหนัก (Weight) และค่าเบี่ยงเบนเริ่มต้น (Bias) ค่าผลลัพธ์ที่ได้จากค่าเริ่มต้นจะถูกนำมาเปรียบเทียบกับผลลัพธ์จริง ค่าที่แตกต่างจะถูกนำมาปรับค่าน้ำหนักและค่าเบี่ยงเบนโดยวิธีลองผิดลองถูก (Trial and Error) จนได้ผลลัพธ์ที่ใกล้เคียงหรือตรงกับผลลัพธ์จริง ค่าน้ำหนักและค่าเบี่ยงเบนสุดท้ายจะถูกนำมาใช้ในการพยากรณ์ผลลัพธ์ที่เกิดขึ้นจากข้อมูลนำเข้ามาใหม่ โดยการคำนวณนี้เริ่มที่ชั้นส่งข้อมูลออกและแพร่ ย้อนกลับไปยังชั้นรับข้อมูลนำเข้า

2.5 ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine: SVM)[6]

หลักการวิเคราะห์ข้อมูลของอัลกอริทึมนี้จะอาศัยพื้นฐานทฤษฎีการเรียนรู้ทางสถิติ เป็นเทคนิคที่ได้รับความนิยมอยากแพร่หลายในงานที่เกี่ยวข้องกับการจำแนกประเภทข้อมูล

เป็นโมเดลที่ใช้การคำนวณทางคณิตศาสตร์ซึ่งจัดอยู่ในกลุ่มเดียวกับ Neural Network แต่ SVM มีข้อดีกว่าตรงที่ไม่ค่อยเกิดปัญหา Overfitting มากเหมือนกับ Neural Network

2.6 Naïve Bayes[6]

ใช้การคำนวณความน่าจะเป็นซึ่งถูกใช้ในการทำนายผล จัดเป็นเทคนิคในการแก้ปัญหาแบบ Classification ที่สามารถคาดการณ์ผลลัพธ์ได้และสามารถอธิบายได้ โดยจะทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปร เพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์

การเรียนรู้เบย์อย่างง่ายเป็นวิธีจำแนกประเภทข้อมูลที่มีประสิทธิภาพวิธีหนึ่งที่การทำงานไม่ซับซ้อนเหมาะกับกรณีของเซตตัวอย่างจำนวนมากและคุณลักษณะ ของตัวอย่างจะไม่ขึ้นต่อกัน

2.7 การถดถอยโลจิสติก (Logistic Regression)[6]

การวิเคราะห์การถดถอยโลจิสติกเป็นการหาความสัมพันธ์ระหว่างตัวแปรตาม และตัวแปรอิสระ โดยที่ตัวแปรตามมีลักษณะเป็นตัวแปรเชิงกลุ่ม เมื่อได้แบบแผนความสัมพันธ์หรือสมการพยากรณ์แล้ว สามารถนำแบบแผนดังกล่าวไปใช้ประมาณค่ากลุ่มได้ โดยอาศัยความน่าจะเป็นของการเกิดเหตุการณ์ การวิเคราะห์การถดถอยโลจิสติกแบ่งออกเป็น 2 ประเภท คือ Binary Logistic Regression ตัวแปรตามเป็นตัวแปรเชิงกลุ่มที่มีค่าได้เพียง 2 ค่า เช่น Y = 1 ถ้าคนไข้มีการติดเชื้อ หรือ Y = 0 ถ้าคนไข้ไม่มีการติดเชื้อ Multinomial Logistic Regression ตัวแปรตามเป็นตัวแปรเชิงกลุ่มที่มีค่ามากกว่า 2 ค่า เช่น Y = 0 หมายถึง ไม่เป็นโรคมะเร็ง

Y = 1 หมายถึง การเป็นมะเร็งขั้นต้น

Y = 4 หมายถึง การเป็นมะเร็งขั้นสุดท้าย

2.8 การสร้างโมเดลสมการโครงสร้าง (Structural Equation Modeling: SEM)

เป็นเทคนิคทางสถิติเทคนิคหนึ่งที่ใช้ในการทดสอบ (testing) และประมาณค่า (estimate) ความสัมพันธ์เชิงเหตผุล (causal relationships)

การสร้างโมเดลสมการโครงสร้างมีวัตุประสงค์ได้ทั้งเพื่อการทดสอบทฤษฎี (theory testing) หรือเพื่อสร้างทฤษฎี (theory building)

- กรณีการทดสอบทฤษฎี (theory testing) สร้างโมเดลด้วยวิธีการเชิงอนุมาน (deductive) หรือการวิจัยเชิงปริมาณ (quantitative research) เริ่มต้นจากการศึกษาทฤษฎีและงานวิจัยเพื่อกำหนดโมเดลสมมติฐานที่แสดงเป็นโมเดลความสัมพันธ์เชิงสาเหตุ (causalmodel) ที่จะได้รับการทดสอบจากข้อมูลที่รวบรวมได้ว่ามีความสอดคล้อง (fit) กันหรือไม่ โดยใช้การวิเคราะห์องค์ประกอบเชิงยืนยัน (confirmatory factor analysis)

- กรณีการสร้างทฤษฎี (theory building) สร้างโมเดลด้วยวิธีการเชิงอุปมาน (inductive) หรือการวิจัยเชิงคุณภาพ (qualitative research) แล้วใช้ข้อมูลประมาณค่าของพารามิเตอร์อิสระ (freeparameters) ซึ่งบ่อยครั้งที่สมมติฐานเบื้องต้นอาจมีการปรับโมเดล ในกรณีเช่นนี้ใช้การวิเคราะห์องค์ประกอบเชิงสำรวจ (exploratory factor analysis)

องค์ประกอบที่สำคัญของโมเดลสมการโครงสร้าง คือ ความสัมพันธ์เชิงสาเหตุ โมเดลสมการโครงสร้าง (structural equation model) ซึ่งแสดงถึงความสัมพันธ์เชิงสาเหตุ (causal relationship) ระหว่างตัวแปรภายนอกและตัวแปรภายใน (หรือระหว่างตัวแปรแฝง) ซึ่งอาจเป็นแบบทางเดียวและแบบเส้นเชิงบวก (recursive and linear additive) หรือแบบสองทางและแบบเส้นเชิงบวก (non- recursive and linear additive) และโมเดลการวัด (measurement model) ซึ่งแสดงถึงความสัมพันธ์ระหว่างตัวแปรแฝงกับตัวแปรสังเกตได้

การยืนยันหรือการทดสอบว่าโมเดลที่สร้างขึ้นมีความสอดคล้องกับข้อมูลเชิงประจักษ์หรือไม่นั้น มีสถิติวัดความสอดคล้อง ดังนี้ เช่น

1) ค่าไค-สแควร์ (chi-square) ที่ไม่มีนัยสำคัญ คือค่า p-value สูงกว่า 0.05

2) ค่าสัดส่วนไค-สแควร์/df มีค่าไม่ควรเกิน 2.00

3) ค่า goodness of fit index: GFI, adjusted goodness of fit index:AGFI, comparative fit index: CFI มีค่าตั้งแต่ 0.90 – 1.00

4) ค่า standardized root mean squared residual: standardized RMR, root mean square of error approximation: RMSEA มีค่าต่ำกว่า 0.05

5) ค่า critical n: CN มีค่าเท่ากับ หรือมากกว่า 200 ของกลุ่มตัวอย่าง

6) ค่า largest standardized residual มีค่า -2 ถึง 2

2.9 การจัดกลุ่มข้อมูล (Clustering) ด้วยเทคนิค K-Mean [3]

กำหนดจำนวนกลุ่ม K กลุ่ม และกำหนดศูนย์กลางเริ่มต้นจำนวน K จุด จากนำข้อมูลทั้งหมดจัดเข้ากลุ่มที่มีจุดศูนย์กลางที่อยู่ใกล้ข้อมูลนั้นมากที่สุด โดยคำนวณจากการวัดระยะห่างระหว่างจุดที่น้อยที่สุด แล้วคำนวณจุดศูนย์กลาง K จุดใหม่ โดยหาค่าเฉลี่ยทุกข้อมูลในกลุ่ม ต่อจากนั้นทำซ้ำในข้อ 2. จนกว่าจุดศูนย์กลางจะไม่มีการเปลี่ยนแปลง

2.10 Gaussian Process (GP) [4]

คือระบบ Stochastic Process ซึ่ง ปัจจุบันถูกยอมรับแล้วว่าเป็นเครื่องมือที่มีประสิทธิภาพในการแก้ปัญหา ประเภท Regression, Classification, และ Decision ใน Machine learning สามารถทำงานได้ดีถึงแม้ว่ามี Training Data น้อยและมีประสิทธิภาพ และ Convergence rate ดีกว่า ARMA, NN และ SVR

งานวิจัยที่เกี่ยวข้อง

ชนิดา เสือเปีย และกมล เกียรติเรืองกมลา (2558) [8] ศึกษาเรื่อง “การศึกษาสาเหตุการเกิดโรคเบาหวานชนิดที่ 2 ด้วยเทคนิคการทำเหมืองข้อมูล กรณีศึกษาโรงพยาบาลรัฐแห่งหนึ่ง” โดยใช้เทคนิคการแบ่งกลุ่ม (Cluster Analysis : K-mean) และ (Logistic Regression Analysis) .ในการสร้างโมเดลพยากรณ์หาสาเหตุการเกิดโรคเบาหวาน พบว่าผลที่ได้จากการแบ่งกลุ่มผู้ป่วยออกเป็น 8 กลุ่ม จะถูกนำไปวิเคราะห์การถดถอยโลจิสติกทวิ เพื่อสามารถอธิบายถึงพฤติกรรมของการเป็นโรคในปัจจุบันที่มีผลไปสู่โรคเบาหวานชนิดที่ 2

บรรจบ ตลกุล,จารี ทองคำ,วาทินี สุขมาก (2014) [9] ศึกษาเรื่อง “การสร้างแบบจำลองเพื่อพยากรณ์การเกิดแผลที่เท้าของ

ผู้ป่วยโรคเบาหวานโดยใช้เทคนิคเหมืองข้อมูล” โดยใช้เทคนิคโครงสร้างประสาทเทียม (ANN) ซัพพอร์ตเวคเตอร์แมชชีน (SVM) และ Decision Tree (C4.5) ผลที่ได้พบว่าเทคนิค Decision Tree (C4.5) มีค่าความแม่นยำสูงสุดที่ 75.90%

กิตติพล วิแสง,สิรภัทร เชี่ยวชาญวัฒนา และคำรณ สุนัติ (2009)[10]ศึกษาเรื่อง“การวิเคราะห์ปัจจัยเสี่ยงของโรคเบาหวาน” โดยใช้เทคนิค ตัวแบบโครงข่ายประสาทเทียมแบบ BPNN ตัวแบบโครงข่ายประสาทเทียม แบบ RBFN และตัวแบบ Naïve Bayes ผลที่ได้พบว่า เทคนิคโครงข่ายประสาทเทียม แบบ BPNN ให้ผลลัพธ์ที่มีความถูกต้องมากที่สุด โดยใช้ค่า MAE เท่ากับ 0.0045 ค่า RMSE เท่ากับ 0.0126

Xue-Hui Meng a, Yi-Xiang Huang a, Dong-Ping Rao b, Qiu Zhang a, Qing Liu b[11] ศึกษาเรื่อง “Comparison of three data mining models for predicting diabetes or prediabetes by risk factors” โดยใช้เทคนิค Logistic Regression, Decision Tree และ Artificial Neural Network: ANN ผลที่ได้พบว่าเทคนิค Decision Tree ให้ผลลัพธ์ที่มีความแม่นยำมากที่สุด ที่ 77.87%

Asma A. AlJarullah[12] (2011) ศึกษาเรื่อง “Decision Tree Discovery for the Diagnosis of Type II Diabetes” โดยใช้เทคนิค Decision Tree (J48) โดยผลที่ได้พบว่า Decision Tree (J48)มีประสิทธิภาพค่าความถูกต้องสูงที่สุดมีค่าเท่ากับ 78.1768%

Sankaranarayanan.S , Dr Pramananda Perumal.T(2014)[13] ศึกษาเรื่อง “A Predictive Approach for Diabetes Mellitus Disease through Data Mining Technologies ” โดยใช้เทคนิค Decision Tree ในการสร้างตัวแบบพยากรณ์

Eleni I. Georga, Vasilios C. Protopappas, Stavroula G. Mougiakakou (2013)[14] ศึกษาเรื่อง“Short-term vs. Long-term Analysis of Diabetes Data: Application of Machine Learning and Data Mining Techniques” โดยใช้เทคนิค **Gaussian Process (GP) และ SVR .ในการสร้างตัวแบบพยากรณ์**

จารุพร สฤษฎิ์สุข,สมชาย ปราการเจริญ(2011) [15] ศึกษาเรื่อง “การพยากรณ์ความเสี่ยงในการเป็นโรคเบาหวานโดยวิธีแบบจำลองสมการโครงสร้าง” โดยใช้เทคนิคการสร้างโมเดลสมการโครงสร้าง (Structural Equation Modeling: SEM) ผลที่ได้พบว่า มีความถูกต้องที่ 56.33%

3. วิธีดำเนินวิจัย

ปัจจุบันทางด้านสาธารณสุขทั่วโลกต่างให้ความสำคัญกับการจัดการโรคเรื้อรังเพิ่มมากขึ้น โดยโรคเบาหวานเป็นอีก 1 โรคเรื้อรังที่มีผู้ป่วยเพิ่มมากขึ้น โรคเบาหวานเป็นโรคที่ติดอันดับ 1 ใน 5 โรคของเอเชีย และติดอันดับ 1 ใน 10 ของโรคที่คุกคามคนไทยมากที่สุด ในประเทศไทยพบว่ามีผู้เสียชีวิตจากโรคเบาหวานเฉลี่ยสูงถึงวันละ 27 ราย อีกทั้งยังเป็นโรคที่มีสถิติแนวโน้มของผู้ป่วยเพิ่มสูงขึ้นอย่างต่อเนื่องทุกๆ ปีอีกด้วย

ดังนั้นจึงมีแนวคิดในการนำเทคโนโลยีการทำเหมืองข้อมูลมาใช้ในการพยากรณ์ โดยได้

ทำการศึกษาเทคนิคต่างๆ เพื่อพิจารณาเทคนิคที่เหมาะสมในการนำมาวิเคราะห์ข้อมูลการพยากรณ์ความเสี่ยงของการเกิดโรคเบาหวาน

งานวิจัยนี้ได้นำเทคนิคการทำเหมืองข้อมูลมาใช้ เพื่อใช้ในการวิเคราะห์การพยากรณ์ความเสี่ยงของการเกิดโรคเบาหวาน โดยวิเคราะห์จากผลการวิจัยของงานวิจัยที่ได้ทำการศึกษา

4. ผลการดำเนินการวิจัย

-

5. สรุป

งานวิจัยนี้เป็นการศึกษาเทคนิคการทำเหมืองข้อมูลเพื่อพยากรณ์ความเสี่ยงการเกิดโรคเบาหวาน โดยผู้วิจัยได้ทำการศึกษางานวิจัยที่เกี่ยวข้อง เพื่อค้นหาเทคนิคที่เหมาะสมในการพยากรณ์ และจากการศึกษาพบว่าเทคนิคต้นไม้ตัดสินใจ (Decision Tree) และโครงข่ายประสาทเทียมแบบแพร่กลับ (Backpropagation Neural Network: BPNN) มีค่าความถูกต้องและมีความเหมาะสมในการนำไปใช้สร้างแบบจำลองพยากรณ์ความเสี่ยงของการเกิดโรคเบาหวาน

เอกสารอ้างอิง

[1] นุชรี อาบสุวรรณ,นิตยา พันธุเวทย์. 2558."**ประเด็นรณรงค์วันเบาหวานโลก ปี 2557**".(ปี งบประมาณ 2558)

[2] Automated, M. (ม.ป.ป.). **เหมืองข้อมูล (data mining)**. สืบค้นเมื่อ 16 พฤศจิกายน 2016, จาก http://www.autosoft.in.th/data-warehouse/

[3] Gintamaz. (2555). **Data Mining: เทคนิคต่าง ๆ ของ Data Mining**. สืบค้นเมื่อ 16 พฤศจิกายน 2016, จาก http://llem0nz.blogspot.com/2012/03/data-mining_8699.html

[4] E. I. Georga, V. C. Protopappas, D. Ardigo, M. Marina, I. Zavaroni, D. Polyzos, and D. I. Fotiadis, "**Multivariate prediction of subcutaneous glucose concentration in type 1 diabetes patients based on support vector regression**", IEEE J Biomed Health Inform., vol. 17, pp. 71-81, 2013.

[5] บุญเสริมกิจศิริกุล. ปัญญาประดิษฐ์, **ภาควิชาวิศวกรรมคอมพิวเตอร์,คณะวิศวกรรมศาสตร์จุฬาลงกรณ์มหาวิทยาลัย**, หน้า 169 –181,2546

[6] Linoff GS, Berry MJA. **Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management**. John Wiley & Sons; 2011. 890 p

[7] รายงานสถิติสาธารณสุข สำนักนโยบายและกลยุทธ์ กระทรวงสาธารณสุข ปี พ.ศ.(2556).**จำนวนและอัตราผู้เสียชีวิตจากโรคเบาหวาน**.

[8] ชนิดา เสือเปีย และกมล เกียรติเรืองกมลา. (2558).**การศึกษาสาเหตุการเกิดโรคเบาหวานชนิดที่ 2 ด้วยเทคนิคการทำเหมืองข้อมูล กรณีศึกษาโรงพยาบาลรัฐแห่งหนึ่ง**. จาก www.ncteched.org

[`9] บรรจบ ดลกุล. (2014). **การสร้างแบบจำลองเพื่อพยากรณ์การเกิดแผลที่เท้าของผู้ป่วยโรคเบาหวานโดยใช้เทคนิคเหมืองข้อมูล.**

[`10] กิตติพล วิแสง,สิรภทัร เชียวชาญวัฒนา และคำรณ สุนัติ. (2009). **การวิเคราะห์ปัจจัยเสียงของโรคเบาหวาน**.

[`11] Xue-Hui Meng a, Yi-Xiang Huang a, Dong-Ping Rao b,Qiu Zhang a, Qing Liu b. **Comparison of three data mining models for predicting diabetes or prediabetes by risk factors.** The Kaohsiung Journal of Medical SciencesVolume 29, Issue 2

[`12] Asma A. AlJarullah. **Decision Tree Discovery for the Diagnosis of Type II Diabetes.** International Conference on Innovations in Information Technology (2011)

[`13] Sankaranarayanan.S , Dr Pramananda Perumal.T. **A Predictive Approach for Diabetes Mellitus Disease through Data Mining Technologies.** World Congress on Computing and Communication Technologies 2014

[`14] Eleni I. Georga, Vasilios C. Protopappas, Stavroula G. Mougiakakou. **Short-term vs. Long-term Analysis of Diabetes Data: Application of Machine Learning and Data Mining Techniques.** 2013

[`15] จารุพร สฤษฎ์สุข,สมชาย **ปราการการพยากรณ์ความเสี่ยงในการเป็นโรคเบาหวานโดยวิธีแบบจำลองสมการโครงสร้าง.** การประชุมทางวิชาการด้านเทคโนโลยีคอมพิวเตอร์และระบบสารสนเทศประยุกต์ระดับชาติ ครั้งที่ 2 2011