การเปรียบเทียบประสิทธิภาพของเทคนิคเหมืองข้อมูลสำหรับพยากรณ์การเกิดโรค
The efficiency comparison of data mining techniques for patient incidence

อุกฤษฏ์ ศรีสุข[1*], จารี ทองคำ[1]
Ukrit Srisuk[1*], Jaree Thongkam[1]

Received: 8 November 2020 ; Revised: 15 February 2021 ; Accepted: 1 March 2021

บทคัดย่อ
งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาประสิทธิภาพของเทคนิคเหมืองข้อมูลในข้อมูลที่หลากหลาย ข้อมูลในงานวิจัยนี้ประกอบด้วยข้อมูลผู้ป่วยโรคมะเร็งเต้านม ผู้ป่วยโรคเบาหวาน และผู้ป่วยโรคไฮเปอร์ไทรอยด์ โดยข้อมูลทั้งหมดถูกรวบรวมมาจากฐานข้อมูล UCI จำนวนทั้งหมด 3 ชุดข้อมูล ในงานวิจัยนี้ได้นำเอาเทคนิค Machine Learning มาใช้กับการทำเหมืองข้อมูล 5 เทคนิค ได้แก่ Decision Tree C4.5, Naïve Bayes, Neural Networks, Random Forest, Deep Learning มาทำการสร้างแบบจำลองเพื่อการพยากรณ์การเกิดโรค โรคมะเร็งเต้านม โรคเบาหวาน และโรคไฮโปไทรอยด์ ในการวัดประสิทธิภาพ 10-fold cross validation ได้ถูกนำมาใช้ในการแบ่งข้อมูลออกเป็นกลุ่มฝึกสอน และ กลุ่มทดสอบ ค่าความถูกต้อง ค่าความไว และค่าจำเพาะ ได้ถูกนำมาใช้ในการเปรียบเทียบประสิทธิภาพการพยากรณ์ของแต่ละแบบจำลอง จากการทดลองพบว่า เทคนิค Decision Tree C4.5 เป็นเทคนิคที่ดีที่สุดในการสร้างแบบจำลองในการพยากรณ์โรคไฮโปไทรอยด์ โดยให้ค่าความถูกต้อง 99.86 % ค่าความไว 99.85 % และค่าจำเพาะ 100 %

คำสำคัญ: เหมืองข้อมูล มะเร็งเต้านม เบาหวาน ไฮโปไทรอยด์

Abstract
This research aims to study the performance of data mining techniques in medical datasets. The data in this research contaita information of patients with breast cancer, diabetics and patients with hyperthyroidism. All datasets were collected from UCI databasee Mmachine learning, in particular Decision Tree C4.5, Naïve Bayes, Neural Networks, Random Forest and Deep Learning techniques were used to create the models of disease Breast cancer, diabetes and hypothyroidism prediction models. In order to measure the performance of prediction models, 10-fold cross validation was utilized to divide the data into training and testing sets. Accuracy, sensitivity and specificity of the prediction models were used to compare the prediction performance of each model. The experimental results showed that the Decision Tree C4.5 technique was the best technique in modeling the prognosis of hypothyroidism. It provided 99.86 % accuracy, 99.85 % sensitivity and 100 % specificity.

Keywords: Data mining, Breast cancer, Diabetics, Hypothyroid

[1] กลุ่มสารสนเทศเชิงประยุกต์ ภาควิชาเทคโนโลยีสารสนเทศ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม
[1] Applied Informatics Group, Information Technology, Faculty of Informatics, Mahasarakham University
* Corresponding author: Ukrit Srisuk , E-mail: ozillmammoth39@gmail.com

บทนำ

มะเร็งเต้านม เป็นมะเร็งที่พบมากที่สุดเป็นอันดับ 1 ของผู้หญิงไทย และเป็นสาเหตุของการเสียชีวิตอันดับต้นๆ ในผู้หญิง แนวโน้มคนไทยป่วยเป็นโรคมะเร็งสูงขึ้นทุกปี แต่อัตราการเป็นโรคน้อยกว่าประเทศทางตะวันตก หญิงไทยมีอัตราการพบมะเร็งประมาณ 40 คน ในสตรีวัยเจริญพันธุ์ 100,000 คน ซึ่งถ้าเทียบกับประเทศตะวันตกพบมะเร็งเต้านมได้มากกว่า 100 คน ในสตรีวัยเจริญพันธุ์ 100,000 คน ในผู้ชายพบมะเร็งเต้านมได้เช่นกัน แต่ไม่บ่อยนัก โดยมีอุบัติการณ์ของโรคนี้น้อยกว่าผู้หญิงเกือบ 100 เท่า (ณัฏฐพร นันทิวัฒนา, 2563) ส่วนโรคเบาหวาน เกิดจากเซลล์ร่างกายมีความผิดปกติในขบวนการเปลี่ยนน้ำตาลในเลือดให้เป็นพลังงาน โดยขบวนการนี้เกี่ยวข้องกับอินซูลินซึ่งเป็นฮอร์โมนที่สร้างจากตับอ่อนเพื่อใช้ควบคุมระดับน้ำตาลในเลือด เมื่อน้ำตาลไม่ได้ถูกใช้จึงทำให้ระดับน้ำตาลในเลือดสูงขึ้นกว่าระดับปกติ (พิมพ์ใจ อันทานนท์, 2563) ยิ่งไปกว่านั้นโรคไฮเปอร์ไทรอยด์ หมายถึง ภาวะที่ต่อมไทรอยด์ มีการหลั่งฮอร์โมนไทรอยด์ออกมามากเกินไป กระตุ้นให้อวัยวะทั่วร่างกายมีการเผาผลาญสูงกว่าปกติ เป็นสาเหตุทำให้เกิดอาการเจ็บป่วยๆ ต่างขึ้นตามมา เช่น เหนื่อยง่าย ใจสั่น ขี้ร้อนง่าย เหงื่อออกมาก หงุดหงิด นอนไม่หลับ น้ำหนักตัวลดลงอย่างรวดเร็วแบบผิดปกติ เป็นต้น สาเหตุของไทรอยด์เป็นพิษเกิดจากการที่ต่อมไทรอยด์ทำงานมากผิดปกติ จนทำให้ร่างกายมีปริมาณของของฮอร์โมนไทรอยด์มากเกินความต้องการของร่างกายและมีสภาวะเป็นพิษจนส่งผลต่อร่างกายในด้านต่างๆ (เมดไทย, 2563)

การทำเหมืองข้อมูล (Data Mining) (Schuh *et al.*, 2020) คือ กระบวนการวิเคราะห์ข้อมูล เพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้นๆ ในปัจจุบันการทำเหมืองข้อมูลได้ถูกนำไปประยุกต์ใช้ในงานหลายประเภท เช่น การพยากรณ์พันธุ์ต้นไม้ การพยากรณ์ผู้ใช้บัตรเครดิต รวมถึงการพยากรณ์ผู้ป่วยเพื่อพยากรณ์การณ์การอุบัติของโรคต่างๆ เช่น โรคมะเร็งเต้านม โรคเบาหวาน และโรคอื่นๆ เป็นต้น มีนักวิจัยหลายท่านที่ได้ทำการพยากรณ์โรคมะเร็งเต้านม เช่น Fan, Zhu และ Yin (Qi *et al.*, 2010) ได้ศึกษาการทำนายการกลับเป็นซ้ำของมะเร็งเต้านม ด้วยเทคนิค Decision Tree C4.5, CHAID, QUEST, CART, ANN พบว่า เทคนิค Decision Tree C4.5 มีประสิทธิภาพที่ดีที่สุดที่ 71.17% ส่วน Balpande และ Wajgi (Balpande & Wajgi, 2017) ได้ศึกษาการคาดคะเนและการประมาณความรุนแรงของโรคเบาหวาน โดยใช้เทคนิคการทำเหมืองข้อมูล ด้วยเทคนิค CHAID, Naïve Bayes, K-Nearest, Decision Tree ปัจจัยที่มีผลต่อการคาดคะเน คือ อายุ, เพศ และ ดัชนีมวลกาย พบว่า เทคนิค Decision Tree มีประสิทธิภาพที่ดีที่สุด ที่ 72% จะเห็นได้ว่าแต่ละเทคนิคมีประสิทธิภาพที่ไม่แน่นอน ในข้อมูลที่แตกต่างกัน

ดังนั้นงานวิจัยนี้ผู้วิจัยมีความสนใจที่จะศึกษาประสิทธิภาพของเทคนิค Decision Tree C4.5, Naïve Bayes, Neural Networks, Random Forest และ Deep Learning ในการสร้างแบบจำลองเพื่อพยากรณ์การเกิดโรค มะเร็งเต้านม โรคเบาหวาน และโรคไฮโปไทรอยด์ โดยวัดประสิทธิภาพด้วยเทคนิค 10-fold cross validation และแสดงค่าความถูกต้อง ค่าความไว และ ค่าจำเพาะ

วิธีดำเนินการวิจัย

ในการดำเนินการวิจัยนี้ได้ใช้กระบวนการในการทำเหมืองข้อมูลซึ่งมี 4 ขั้นตอนมาใช้ คือ การเตรียมข้อมูล การคัดเลือกตัวแปร การสร้างแบบจำลอง และการวัดประสิทธิภาพของแบบจำลอง

1. การเตรียมข้อมูล

การเตรียมข้อมูลในงานวิจัยนี้ ข้อมูลประกอบด้วยชุดข้อมูลที่มีชนิดตัวแปรที่แตกต่างกันจำนวน 3 ชุดข้อมูล คือ ชุดที่ 1 คือโรคมะเร็งเต้านม มีชนิดตัวแปรเป็น Nominal ทั้งหมด จำนวน 10 ตัวแปร ชุดที่ 2 คือโรคเบาหวาน มีชนิดตัวแปรเป็น Numeric ทั้งหมด จำนวน 9 ตัวแปร และชุดที่ 3 โรคไฮโปไทรอยด์ มีชนิดตัวแปรเป็น Nominal และ Numeric จำนวน 30 ตัวแปร จากฐานข้อมูล UCI

ชุดที่ 1 ตัวแปรที่ใช้ในงานวิจัยโรคมะเร็งเต้านม

1. อายุ (age) มีชนิดข้อมูลเป็น Nominal

2. สตรีวัยหมดประจำเดือน (menopause) มีชนิดข้อมูลเป็น Nominal

3. ขนาดของเนื้องอก (tumor-size) มีชนิดข้อมูลเป็น Nominal

4. ช่วงของมะเร็ง (inv-nodes) มีชนิดข้อมูลเป็น Nominal

5. การกระจายตัวของมะเร็ง (node-caps) มีชนิดข้อมูล Nominal

6. ระดับความร้ายแรง (deg-malig) มีชนิดข้อมูลเป็น Nominal

7. ตำแหน่งของมะเร็ง (breast) มีชนิดข้อมูลเป็น Nominal

8. ตำแหน่งของมะเร็ง (breast-quad) มีชนิดข้อมูลเป็น Nominal

9. การฉายรังสี (Irradiat) มีชนิดข้อมูลเป็น Nominal

10. เหตุการณ์ (class) มีชนิดข้อมูลเป็น Nominal

ชุดที่ 2 ตัวแปรที่ใช้ในงานวิจัยโรคเบาหวาน

1. จำนวนครั้งที่ตั้งครรภ์ (preg) มีชนิดข้อมูลเป็น

Numeric Min=0, Max=17, Mean=3.845, SD.=3.37

2. ระดับน้ำตาลในเลือด (plas) มีชนิดข้อมูลเป็น Numeric Min=0, Max=199, Mean=120.895, SD.=31.973

3. ความดันโลหิต (pres) มีชนิดข้อมูลเป็น Numeric Min=0, Max=122, Mean=69.105, SD.=19.356

4. ความหนาของไขมันใต้ผิวหนังกล้ามแขนส่วนหลัง (skin) มีชนิดข้อมูลเป็น Numeric Min=0, Max=99, Mean=20.536, SD.=15.952

5. ปริมาณอินซูลินที่ได้รับภายใน 2 ชั่วโมง (insu) มีชนิดข้อมูลเป็น Numeric Min=0, Max=846, Mean=79.799, SD.=115.244

6. ดัชนีมวลกาย (mass) มีชนิดข้อมูลเป็น Numeric Min=0, Max=67.1, Mean=31.993, SD.=7.884

7. การติดต่อทางสายเลือด (pedi) มีชนิดข้อมูลเป็น Numeric Min=0.078, Max=2.42, Mean=0.472, SD.=0.331

8. อายุ (age) มีชนิดข้อมูลเป็น Numeric Min=21, Max=81, Mean=33.241, SD.=11.76

9. ตัวแปรคลาส (class) มีชนิดข้อมูลเป็น Nominal

ชุดที่ 3 ตัวแปรที่ใช้ในงานวิจัยโรคไฮโปไทรอยด์

1. อายุ (age) มีชนิดข้อมูลเป็น Numeric Min=1, Max=455, Mean=51.783, SD.=20.124

2. เพศ (sex) มีชนิดข้อมูลเป็น Nominal

3. ฮอร์โมนที่หลั่งออกมาจากต่อมไทรอยด์ (on thyroxine) มีชนิดข้อมูลเป็น Nominal

4. คำถามของฮอร์โมนที่หลั่งออกมาจากต่อมไทรอยด์ (query on thyroxine) มีชนิดข้อมูลเป็น Nominal

5. การใช้ยาต้านไทรอยด์ (on antithyroid medicate) มีชนิดข้อมูลเป็น Nominal

6. อาการไข้ (sick) มีชนิดข้อมูลเป็น Nominal

7. การตั้งครรภ์ (pregnant) มีชนิดข้อมูลเป็น Nominal

8. การตัดต่อมไทรอยด์บางส่วน (thyroid surgery) มีชนิดข้อมูลเป็น Nominal

9. การรักษาด้วยแร่ไอโอดีน 131 (I131 treatment) มีชนิดข้อมูลเป็น Nominal

10. โรคไทรอยด์ชนิดอ้วน (query hypothyroid) มีชนิดข้อมูลเป็น Nominal

11. ภาวะที่ต่อมไทรอยด์สร้างและปล่อยฮอร์โมนออกมามากเกินไป (query hyperthyroid) มีชนิดข้อมูลเป็น Nominal

12. การใช้ยาทางจิตเวชที่ใช้รักษาอาการคลุ้มคลั่ง (lithium) มีชนิดข้อมูลเป็น Nominal

13. ภาวะที่มีการโตขึ้นของต่อมไทรอยด์ (goitre) มีชนิดข้อมูลเป็น Nominal

14. เนื้องอก (tumor) มีชนิดข้อมูลเป็น Nominal

15. ภาวะที่ร่างกายขาดฮอร์โมนที่สร้างจากต่อมใต้สมอง (hypopituitary) มีชนิดข้อมูลเป็น Nominal

16. อัตราซึมพจร (psych) มีชนิดข้อมูลเป็น Nominal

17. ค่าของฮอร์โมนกระตุ้นต่อมไทรอยด์ (TSH measured) มีชนิดข้อมูลเป็น Nominal

18. ฮอร์โมนกระตุ้นต่อมไทรอยด์ (TSH) มีชนิดข้อมูลเป็น Numeric Min=0.005, Max=145, Mean=2.596, SD.=6.27

19. ค่าของไทรอยด์ฮอร์โมนที่ได้จากสองแหล่ง (T3 measured) มีชนิดข้อมูลเป็น Nominal

20. ไทรอยด์ฮอร์โมนที่ได้จากสองแหล่ง (T3) มีชนิดข้อมูลเป็น Numeric Min=0.05, Max=10.6, Mean=2.045, SD.=0.812

21. ค่าของฮอร์โมนทั้งส่วนที่จับกับโปรตีน (TT4 measured) มีชนิดข้อมูลเป็น Nominal

22. ฮอร์โมนทั้งส่วนที่จับกับโปรตีน (TT4) มีชนิดข้อมูลเป็น Numeric Min=19, Max=430, Mean=110.344, SD.=33.692

23. ค่าของการใช้ไทรอกซิน (T4U measured) มีชนิดข้อมูลเป็น Nominal

24. การใช้ไทรอกซิน (T4U) มีชนิดข้อมูลเป็น Numeric Min=0.25, Max=2.32, Mean=0.994, SD.=0.196

25. ค่าของฮอร์โมนไทรอกซินชนิดอิสระ (FTI measured) มีชนิดข้อมูลเป็น Nominal

26. ฮอร์โมนไทรอกซินชนิดอิสระ (FTI) มีชนิดข้อมูลเป็น Numeric Min=17, Max=395, Mean=112.522, SD.=30.871

27. ค่าของโปรตีนที่สร้างจากตับ (TBG measured) มีชนิดข้อมูลเป็น Nominal

28. โปรตีนที่สร้างจากตับ (TBG) มีชนิดข้อมูลเป็น Numeric

29. แหล่งอ้างอิง (referral source) มีชนิดข้อมูลเป็น Nominal

30. ตัวแปรคลาส (Class) มีชนิดข้อมูลเป็น Nominal

2. การคัดเลือกตัวแปร

การคัดเลือกตัวแปรเป็นวิธีการในเลือกตัวแปรตามที่มีความสัมพันธ์กับตัวแปรตาม และยังช่วยในการลดจำนวนตัวต้นที่ไม่เกี่ยวข้องกับตัวแปรตามออกจากชุดข้อมูลที่จะนำไปสร้างแบบจำลองได้ โดยในงานวิจัยนี้จะใช้เทคนิค Chi-Square

Chi-Square เป็นการประเมินค่าของคุณลักษณะโดยการใช้การคำนวณค่า Chi-Square ทางสถิติเพื่อศึกษาว่าการแจกแจงความถี่ของตัวแปรคุณลักษณะเป็นไปตามรูปแบบที่กำหนดไว้ หรือไม่ ดังสมการ

$$x^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

โดยที่

O1 , O2 … On เป็นความถี่ของตัวแปรที่ได้จากการศึกษา

E1, E2 … En เป็นความถี่ที่คาดหวัง (หรือความถี่ที่ควรจะเป็น)

3. การสร้างแบบจำลอง

การสร้างแบบจำลองเป็นขั้นตอนที่นำเอาเทคนิคในMachine Learning มาใช้มาทำการสร้างแบบจำลองเพื่อใช้ในการพยากรณ์ ด้วยโปรแกรม Weka โดยใช้เทคนิคดังต่อไปนี้

เทคนิคต้นไม้ตัดสินใจ C4.5 (Decision Tree C4.5) (ชณิดาภา บุญประสม, 2563) เป็นอัลกอริทึม ที่พัฒนามาจากอัลกอริทึม ID3 เป็นอัลกอริทึมในการจำแนกประเภทข้อมูล ใช้หลักการสร้างต้นไม้โดยคัดเลือกคุณลักษณะที่สำคัญที่สุดมาเป็นโหนดราก (Root Node) โดยใช้ค่า Gain Ratio ที่สูงที่สุดเป็นโหนดราก และโหนดถัดไป และต้องหาค่า Entropy, Information Gain และ Split Information

เทคนิคการพยากรณ์ข้อมูลแบบจำลองเบย์ (Naïve Bayes) (Kaur *et al.*,d.) เป็นการพยากรณ์ประเภทโดยใช้กฎของเบย์หรือเป็นการพยากรณ์ประเภทโดยใช้หลักสถิติในการพยากรณ์ความน่าจะเป็น

เทคนิคโครงข่ายประสาทเทียม (Artificial Neural Networks) (ทิพย์หทัย ทองธรรมชาติ, 2560) เป็นศาสตร์แขนงหนึ่งทางด้านปัญญาประดิษฐ์ (Artificial Intelligence : AI) ที่สามารถนำไปประยุกต์ใช้กับงานหลายด้านได้อย่างมีประสิทธิภาพ หลักการสำคัญของโครงข่ายประสาทเทียม คือ ความพยายามที่จะลอกเลียนแบบการทำงานของเซลล์ประสาทในสมองมนุษย์เพื่อทำงานได้อย่างมีประสิทธิภาพ โดยที่โครงข่ายประสาทเทียมแบบหลายชั้น (multilayer) มีลักษณะเช่นเดียวกับโครงข่ายประสาทเทียมแบบชั้นเดียว แต่จะมีชั้นแอบแฝง (hidden) เพิ่มขึ้น โดยอยู่ส่วนกลางระหว่างชั้นนำข้อมูลป้อนเข้าและชั้นส่งขอมูลออกทั้งนี้ชั้นแอบแฝงอาจมีมากกว่า 1 ชั้น

เทคนิคต้นไม้ป่าสุ่ม (Random Forest) (Rastgou *et al.*, 2020) เป็น Model ประเภทหนึ่งของ Machine Learning ถูกพัฒนาขึ้นจาก Decision Tree ต่างกันที่ Random Forest เป็นการเพิ่มจำนวน Tree เป็น Tree หลายๆ ต้น ทำให้ประสิทธิภาพในการทำงานสูงขึ้น แม่นยำมากขึ้น ซึ่งโมเดล Random Forest เป็นโมเดลที่ได้รับความนิยมเป็นอย่างมากในการใช้ Machine Learning

เทคนิคการเรียนรู้เชิงลึก (Deep Learning) (Dogan & Birant, 2021) เป็นส่วนหนึ่งของ Machine Learning บนพื้นฐานของโครงข่ายประสาทเทียมและการเรียนเชิงคุณลักษณะ การเรียนรู้สามารถเป็นได้ทั้งแบบการเรียนรู้แบบมีผู้สอน การเรียนรู้แบบกึ่งมีผู้สอน และการเรียนรู้แบบไม่มีผู้สอน โดยงานวิจัยนี้เป็นการเรียนรู้ Machine Learning แบบมีผู้สอน (Supervised learning)

4. การวัดประสิทธิภาพของแบบจำลอง

เมื่อทำการสร้างแบบจำลองเสร็จแล้วนำแบบบจำลองมาทดสอบประเมินประสิทธิภาพด้วยวิธีการของ 10-fold cross validation โดยการแบ่งข้อมูลออกเป็น 10 กลุ่มเท่าๆ กันและทำการเปรียบเทียบค่าด้วยการพยากรณ์กลุ่มข้อมูล คือ ค่าความถูกต้อง (Accuracy) ค่าความไว (Sensitivity) และค่าจำเพาะ (Specificity) ดังสมการ

1.ค่าความถูกต้อง (Accuracy) คือ ค่าที่แบบจำลองสามารถพยากรณ์ข้อมูลผู้ป่วยที่เกิดโรค และไม่เกิดโรคได้อย่างถูกต้องต่อข้อมูลทั้งหมด ดังสมการที่ 1

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

2.ค่าความไว (Sensitivity) คือ ค่าที่แบบจำลองสามารถพยากรณ์ข้อมูลผู้ป่วยที่เกิดโรค ได้อย่างถูกต้องต่อผู้ป่วยที่เกิดโรคจริง ดังสมการที่ 2

$$\text{Sensitivity} \ \frac{TP}{TP + FN} \quad (2)$$

3.ค่าจำเพาะ (Specificity) คือ ค่าที่แบบจำลองสามารถพยากรณ์ข้อมูลผู้ป่วยที่ไม่เกิดโรค ได้อย่างถูกต้องต่อผู้ป่วยที่พยากรณ์ว่าเกิดโรค ดังสมการที่ 3

$$\text{Specificity} \quad \frac{TN}{TN + FP} \qquad (3)$$

เมื่อ TP คือ จำนวนข้อมูลที่แบบจำลองพยากรณ์การเกิดโรคได้อย่างถูกต้อง

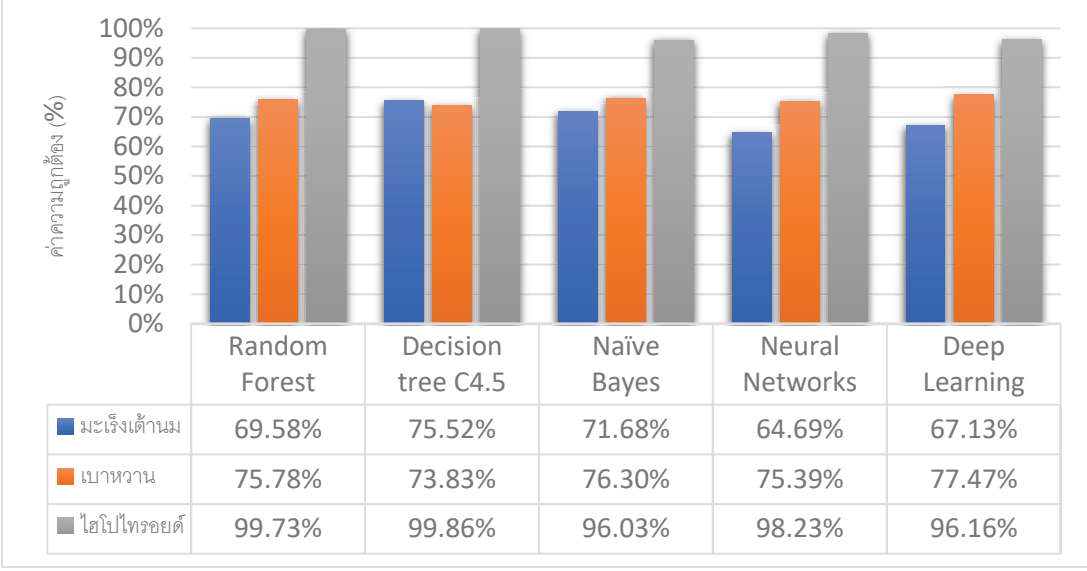
TN คือ จำนวนข้อมูลที่แบบจำลองพยากรณ์การไม่เกิดโรคได้อย่างถูกต้อง

FP คือ จำนวนข้อมูลที่ไม่เกิดโรคแต่แบบจำลองพยากรณ์ว่าเกิดโรค

FN คือ จำนวนข้อมูลที่เกิดโรคแต่แบบจำลองพยากรณ์ว่าไม่เกิดโรค

ผลการวิจัย

งานวิจัยนี้ได้ศึกษาประสิทธิภาพของเทคนิค Machine Learning ของการทำเหมืองข้อมูล ในการสร้างแบบจำลองเพื่อพยากรณ์การเกิดโรค มะเร็งเต้านม โรคเบาหวาน และโรคไฮเปอร์ไทรอยด์ ด้วยเทคนิค Decision Tree C4.5, Naïve Bayes, Neural Networks, Random Forest, และ Deep Learning ผ่านการใช้งานโปรแกรม WEKA เวอร์ชัน 3.9.3 ในการทำการทดลอง ข้อมูลที่ใช้ในงานวิจัยนี้เก็บรวบรวมจากฐานข้อมูล UCI จำนวน 3 ชุดข้อมูล ชุดที่ 1 คือโรคมะเร็งเต้านม มีชนิดตัวแปรเป็น Nominal ทั้งหมด ชุดที่ 2 คือโรคเบาหวาน มีชนิดตัวแปรเป็น Numeric และ Nominal ชุดที่ 3 โรคไฮโปไทรอยด์ มีชนิดตัวแปรเป็น Numeric และ Nominal โดยใช้การทดสอบโมเดลด้วยวิธีการ 10-Fold Cross Validation ในการแบ่งกลุ่มข้อมูลเป็นชุดข้อมูลเรียนรู้ และชุดข้อมูลทดสอบ และวัดประสิทธิภาพของแบบจำลองด้วยค่าความถูกต้อง ค่าความไว และค่าจำเพาะ ดังแสดงใน Figure 1-3

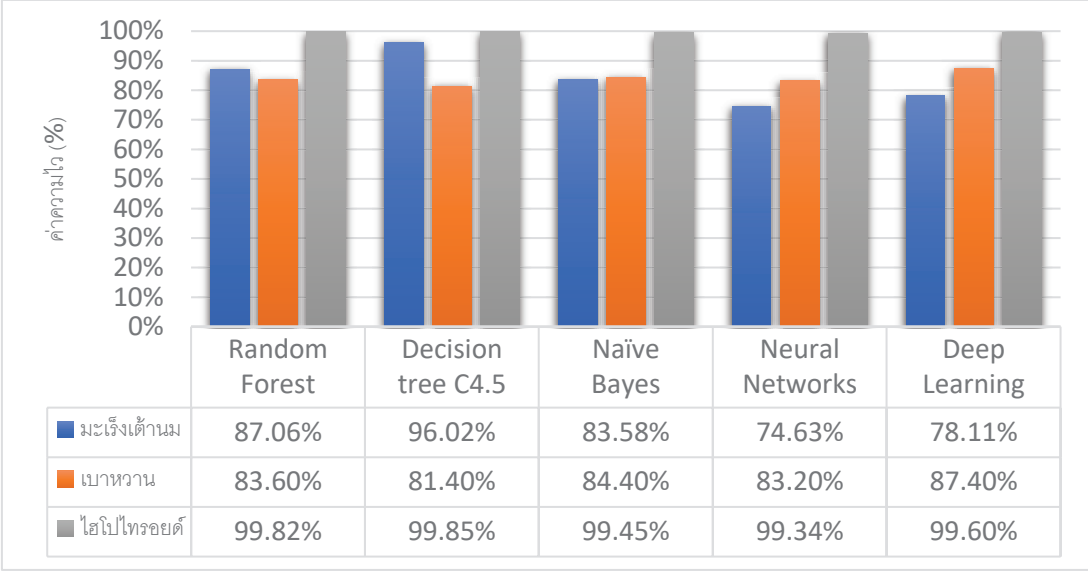


	Random Forest	Decision tree C4.5	Naïve Bayes	Neural Networks	Deep Learning
มะเร็งเต้านม	69.58%	75.52%	71.68%	64.69%	67.13%
เบาหวาน	75.78%	73.83%	76.30%	75.39%	77.47%
ไฮโปไทรอยด์	99.73%	99.86%	96.03%	98.23%	96.16%

Figure 1 Accuracy comparison

Figure 1 การเปรียบเทียบค่าความถูกต้อง (Accuracy) โดยใช้เทคนิค Decision Tree C4.5, เทคนิค Random Forest เทคนิค Naïve Bayes เทคนิค Artificial Neural Networks และเทคนิค Deep Learning ในการพยากรณ์การเกิดโรค ผลปรากฏว่า ในข้อมูลโรคมะเร็งเต้านม ซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Nominal เทคนิค Decision Tree C4.5 สามารถสร้างแบบจำลองที่มีค่าความถูกต้องในการพยากรณ์สูงที่สุดถึง 75.52% และน้อยที่สุดใน เทคนิค Artificial Neural Networks ให้ค่าความถูกต้องที่ 64.69% ในข้อมูลโรคเบาหวาน

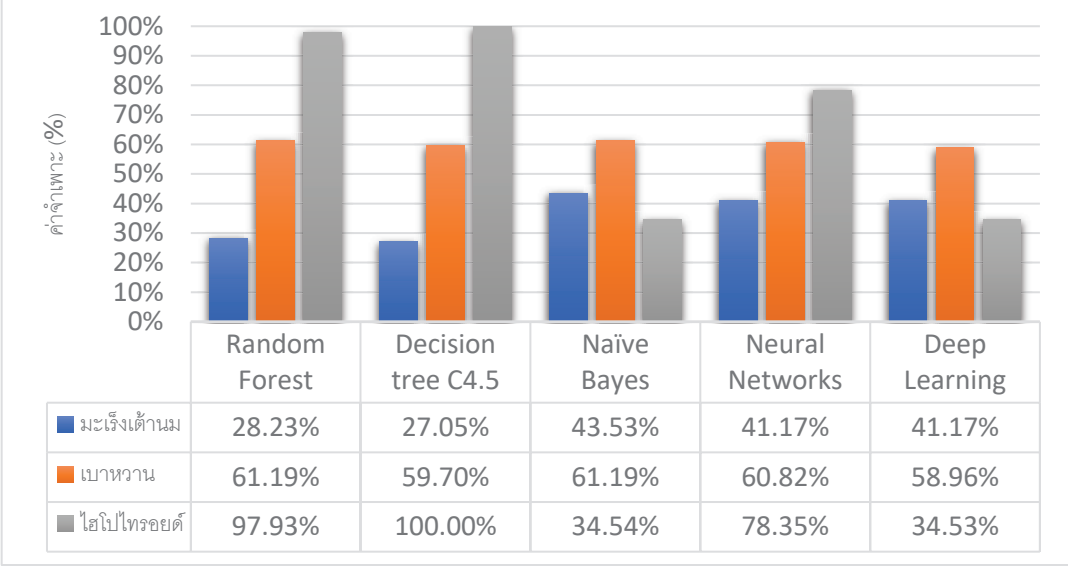
ซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Numeric เทคนิค Deep Learning สามารถสร้างแบบจำลองที่มีค่าความถูกต้องในการพยากรณ์สูงที่สุดถึง 77.47% และน้อยที่สุดในเทคนิค Decision Tree C4.5 ให้ค่าความถูกต้องที่ 73.83% และโรคไฮโปไทรอยด์ ซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Nominal และ Numeric เทคนิค Decision Tree C4.5 สามารถสร้างแบบจำลองที่มีค่าความถูกต้องในการพยากรณ์สูงที่สุดถึง 99.86% และน้อยที่สุดในเทคนิค Naïve Bayes ให้ค่าความถูกต้อง ที่ 96.03%



	Random Forest	Decision tree C4.5	Naïve Bayes	Neural Networks	Deep Learning
มะเร็งเต้านม	87.06%	96.02%	83.58%	74.63%	78.11%
เบาหวาน	83.60%	81.40%	84.40%	83.20%	87.40%
ไฮโปไทรอยด์	99.82%	99.85%	99.45%	99.34%	99.60%

Figure 2 Sensitivity comparison

Figure 2 การเปรียบเทียบค่าความไว (Sensitivity) โดยใช้เทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning ในการพยากรณ์การเกิดโรค ผลปรากฏว่า ในข้อมูลโรคมะเร็งเต้านมซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Nominal เทคนิค Decision Tree C4.5 สามารถสร้างแบบจำลองที่มีค่าความไวในการพยากรณ์สูงที่สุดถึง 96.02% และน้อยที่สุดในเทคนิค Artificial Neural Networks ให้ค่าความไวที่ 74.63% ในข้อมูลโรคเบาหวานซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Numeric เทคนิค Deep Learning สามารถสร้างแบบจำลองที่มีค่าความไวในการพยากรณ์สูงที่สุดถึง 87.40% และน้อยที่สุดในเทคนิค Decision Tree C4.5 ให้ค่าความไวที่ 81.40% และ โรคไฮโปไทรอยด์ ซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Nominal และ Numeric เทคนิค Decision Tree C4.5 สามารถสร้างแบบจำลองที่มีค่าความไวในการพยากรณ์สูงที่สุดถึง 99.85% และน้อยที่สุดในเทคนิค Artificial Neural Networks ให้ค่าความไวที่ 99.34%



	Random Forest	Decision tree C4.5	Naïve Bayes	Neural Networks	Deep Learning
มะเร็งเต้านม	28.23%	27.05%	43.53%	41.17%	41.17%
เบาหวาน	61.19%	59.70%	61.19%	60.82%	58.96%
ไฮโปไทรอยด์	97.93%	100.00%	34.54%	78.35%	34.53%

Figure 3 Specific value comparison

Figure 3 การเปรียบเทียบค่าจำเพาะ (Specificity) โดยใช้เทคนิค Decision Tree C4.5, เทคนิค Random Forest, เทคนิค Naïve Bayes, เทคนิค Artificial Neural Networks และเทคนิค Deep Learning ในการพยากรณ์การเกิดโรค ผลปรากฏว่า ในข้อมูลโรคมะเร็งเต้านมซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Nominal เทคนิค Naïve Bayes ให้ค่าจำเพาะในการพยากรณ์สูงที่สุดถึง 43.53% และน้อยที่สุดในเทคนิค Decision Tree C4.5 ให้ค่าจำเพาะ ที่ 27.05% ในข้อมูลโรคเบาหวาน ซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Numeric เทคนิค Random Forest กับ เทคนิค Naïve Bayes ให้ค่าจำเพาะในการพยากรณ์สูงที่สุดเท่ากันที่ 61.19% และน้อยที่สุดในเทคนิค Deep Learning ให้ค่าจำเพาะที่ 58.96% และ โรคไฮโปไทรอยด์ ซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Nominal และ Numeric เทคนิค Decision Tree C4.5 ให้ค่าจำเพาะในการพยากรณ์สูงที่สุดถึง 100% และน้อยที่สุดในเทคนิค Deep Learning ให้ค่าจำเพาะที่ 34.53%

สรุปผลการศึกษา

งานวิจัยฉบับนี้มีวัตถุประสงค์เพื่อศึกษาประสิทธิภาพของเทคนิคเหมืองข้อมูลในข้อมูลที่หลากหลาย โดยการสร้างแบบจำลองเพื่อพยากรณ์การเกิดโรคมะเร็งเต้านม โรคเบาหวาน และโรคไฮโปไทรอยด์ จากฐานข้อมูล UCI จำนวนทั้งหมด 3 ชุดข้อมูล ด้วยเทคนิค Random Forest เทคนิค Decision Tree C4.5 เทคนิค Naïve Bayes เทคนิค Artificial Neural Networks และเทคนิค Deep Learning จากการทดลองพบว่า ในข้อมูลโรคมะเร็งเต้านมซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Nominal เทคนิค Decision Tree C4.5 สามารถสร้างแบบจำลองที่มีค่าความถูกต้องในการพยากรณ์สูงที่สุดถึง 75.52% ในข้อมูลโรคเบาหวานซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Numeric เทคนิค Deep Learning สามารถสร้างแบบจำลองที่มีค่าความถูกต้องในการพยากรณ์สูงที่สุดถึง 77.47% และ โรคไฮโปไทรอยด์ ซึ่งเป็นข้อมูลที่มีตัวแปรเป็นแบบ Nominal และ Numeric เทคนิค Decision Tree C4.5 สามารถสร้างแบบจำลองที่มีค่าความถูกต้องในการพยากรณ์สูงที่สุดถึง 99.86% จากผลการทดลอง สรุปได้ว่าเทคนิค Decision Tree C4.5 มีความเหมาะสมในการนำสร้างแบบจำลองเพื่อพยากรณ์การเกิดโรคมะเร็งเต้านม โรคเบาหวาน และโรคไฮโปไทรอยด์ เพราะ สามารถจัดการกับข้อมูลที่มีหลาย ๆมิติหรือข้อมูลที่มีหลาย ๆตัวแปรได้ และ ผลการพยากรณ์ข้อมูลมีความถูกต้องค่อนข้างสูง

กิตติกรรมประกาศ

ผู้วิจัยขอขอบคุณเว็บไซต์ UCI ที่ให้ข้อมูลในการทำวิจัยครั้งนี้

เอกสารอ้างอิง

ชณิดาภา บุญประสม. (2563). *การวิเคราะห์การทำนายการลาออกกลางคันของนักศึกษาระดับปริญญาตรีโดยใช้เทคนิควิธีการทำเหมืองข้อมูล*. http://research.fte.kmutnb.ac.th/download.php?filename=620701000056&filepath=20190701155744.pdf.

ณัฏฐพร นันทิวัฒนา. (2563). *มะเร็งเต้านม*. https://www.sikarin.com/content/detail/461/โรคมะเร็งเต้านม-มะเร็งอันดับ-1-ของผู้หญิง.

ทิพย์หทัย ทองธรรมชาติ. (2560). *การคัดเลือกคุณลักษณะเพื่อสร้างโมเดลสำหรับการพยากรณ์ผลสัมฤทธิ์ทางการเรียนด้วยเทคนิคเหมืองข้อมูล*. https://research.kpru.ac.th/sac/fileconference/10912018-05-01.pdf.

พิมพ์ใจ อันทานนท์. (2563). *โรคเบาหวาน*. https://www.dmthai.org/index.php/knowledge/for-normal-person/health-information-and-articles/health-information-and-articles-old-3/846-2019-04-20-01-49-18.

เมดไทย. (2563). *ไทรอยด์เป็นพิษ*. https://medthai.com/ไทรอยด์เป็นพิษ.

Balpande, V.R. and Wajgi, R.D. (2017). Prediction and severity estimation of diabetes using data mining technique. *2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, 576-580.

Dogan, A. and Birant, D. (2021). Machine learning and data mining in manufacturing. *Expert Systems with Applications, 166*, 114060.

Kaur, A.G. M. (n.d.). *A Framework for the Indirect Assessment Tool for Outcome Based Education Using Data Mining*. https://ieeexplore.ieee.org/document/8782336.

Qi, F., Chang-jie, Z. and Liu, Y. (2010). Predicting breast cancer recurrence using data mining techniques. *International Conference on Bioinformatics and Biomedical Technology*, 310-311.

Rastgou, M., Bayat, H., Mansoorizadeh, M. and Gregory, A.S. (2020). Estimating the soil water retention curve: Comparison of multiple nonlinear regression approach and random forest data mining technique. *Computers and Electronics in Agriculture,*174, 105502.

Schuh, G., Prote, J.-P. and Hünnekes, P. (2020). Data mining methods for macro level process planning, *Procedia CIRP*/03/15/ 2021.