การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลการเกิดโรคไวรัสตับอักเสบ

ชุติมา สุขสมัย[1] , ไก้รุ่ง เฮงพระพรหม[1*] และ สุพจน์ เฮงพระพรหม[1]

[1] สาขาวิชาวิทยาการข้อมูล คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครปฐม
*kairung2011.heng@gmail.com

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อ 1) เพื่อศึกษาแนวทางในการจำแนกข้อมูลการเกิดไวรัสตับอักเสบ 2) เพื่อเปรียบเทียบประสิทธิภาพในการจำแนกโรคไวรัสตับเอกเสบด้วยเทคนิคการจำแนกข้อมูล 3 วิธี ได้แก่ วิธีเพื่อนบ้านที่ใกล้ที่สุด ต้นไม้ตัดสินใจ และนาอีฟเบย์ จากการทดลองพบว่า วิธีนาอีฟเบย์และต้นไม้ตัดสินใจให้ประสิทธิภาพที่ดีที่สุด โดยให้ค่าความแม่นยำ 100% ค่าความถูกต้อง 100% และค่าความระลึก 100% ส่วนรองลงมาคือ วิธีเพื่อนบ้านที่ใกล้ที่สุดให้ค่าความแม่นยำ 91% ค่าความถูกต้อง 91% และค่าความระลึก 100%

คำสำคัญ: วิธีเพื่อนบ้านที่ใกล้ที่สุด ต้นไม้ตัดสินใจ นาอีฟเบย์ ชุดข้อมูลโรคไวรัสตับอักเสบ การจำแนกข้อมูล

A Comparison of Data Classification Efficiency for Hepatitis

Chutima Suksamai[1] Kairung Hengpraprohm[1*], and Supojn Hengpraprohm[1]

[1]Program in data science, Faculty of Science and Technology, Nakhon Pathom Rajabhat University
*kairung2011.heng@gmail.com

Abstract

The objectives of this research are: 1) to study the methods for data classification of hepatitis virus occurrence, and 2) to compare the efficiency of hepatitis data classification by using 3 techniques including k Nearest Neighbor, Decision Tree, and Naive Bayes. The results show Naive Bayes and the Decision Tree give the best performance by providing 100% accuracy, 100% accuracy, and 100% recall followed by k Nearest neighbor with 91% accuracy, 91% accuracy, and 100% recall.

Keywords: K-Nearest neighbor, Decision Tree, Naive Bayes, Hepatitis B Virus Infection, Data Classification

1. บทนำ

สถานการณ์ของโรคตับอักเสบทั่วโลกจากข้อมูลขององค์การอนามัยโลก พบผู้ติดเชื้อไวรัสตับอักเสบบีเรื้อรังประมาณ 257 ล้านคนและผู้ติดเชื้อไวรัสตับอักเสบซีเรื้อรังประมาณ 71 ล้านคน สถานการณ์ในประเทศไทย พบผู้ติดเชื้อไวรัสตับอักเสบ บี ประมาณ 2.2-3 ล้านคน และผู้ติดเชื้อไวรัสตับอักเสบซี ประมาณ 8 แสนคน พบมากในผู้ที่มีอายุ 30 ปีขึ้นไป และพบอัตราการเสียชีวิตมีแนวโน้มเพิ่มสูงขึ้น

โรคไวรัสตับอักเสบ บี และซี สามารถติดต่อผ่านทางเลือดและสารคัดหลั่ง การใช้เข็มฉีดยาร่วมกัน การมีเพศสัมพันธ์โดยไม่สวมถุงยางอนามัย สำหรับโรคไวรัสตับอักเสบบี สามารถติดต่อจากแม่สู่ลูกได้ ทั้งนี้ โรคไวรัสตับอักเสบบี มีวัคซีนในการป้องกัน หากมีพฤติกรรมเสี่ยงให้รีบไปตรวจคัดกรองโดยเร็ว สำหรับหญิงตั้งครรภ์ ควรรีบมาฝากครรภ์ และรับการตรวจหาโรคไวรัสตับอักเสบบี เพื่อลดการถ่ายทอดเชื้อจากแม่สู่ลูกได้อย่างมีประสิทธิภาพ ปัจจุบันได้มีการใช้วัคซีนเพื่อสร้างเสริมภูมิคุ้มกันในเด็กแรกเกิดทุกคน เพื่อป้องกันโรคไวรัสตับอักเสบบี แต่สำหรับประชาชนที่เกิดก่อนปี 2535 ควรตรวจหาภูมิคุ้มกันโรคไวรัสตับอักเสบบี หากไม่มีภูมิคุ้มกัน แนะนำให้ฉีดวัคซีนป้องกันโรคไวรัสตับอักเสบบี ให้ครบชุด จำนวน 3 เข็ม ซึ่งจะป้องกันไวรัสตับอักเสบบี และลดจำนวนผู้ติดเชื้อรายใหม่ได้ ส่วนโรคไวรัสตับอักเสบซี ไม่มีวัคซีนในการป้องกัน แต่สามารถรักษาให้หายขาดได้ด้วยการทานยาต้านไวรัสให้ครบ 12 สัปดาห์ ผู้ที่มีพฤติกรรมเสี่ยง ควรเข้ารับบริการตรวจคัดกรองและรักษาโดยเร็ว

การเสียชีวิตเนื่องจากสาเหตุการติดเชื้อไวรัสตับอักเสบเรื้อรัง เช่น ภาวะตับวายเฉียบพลัน ภาวะตับแข็ง และมะเร็งตับ พบมากเป็นอันดับที่ 7 ของประชากรทั่วโลก หรือประมาณ 1.4 ล้านคน ในปี พ.ศ.2556 ซึ่งเป็นตัวเลขที่ใกล้เคียงกับจำนวนผู้เสียชีวิตจากเอดส์และวัณโรค และพบว่าสาเหตุเกือบทั้งหมดเกิดจากการติดเชื้อไวรัสตับอักเสบ บีและ ซีเรื้อรังโรคไวรัสตับอักเสบ บีคาดการณ์ว่าปัจจุบันมีผู้ติดเชื้อทั่วโลกประมาณ 240 ล้านคน โดยจำนวนราว 100 ล้านคนเป็นผู้ป่วยในภูมิภาคเอเชียตะวันออกเฉียงใต้ (SEARO) ซึ่งสูงกว่าร้อยละ 5.6 ของประชากรทั้งหมดของภูมิภาคสำหรับในประเทศไทยพบอัตราความชุกของผู้ติดเชื้อไวรัสตับอักเสบ บีเรื้อรัง ประมาณร้อยละ4-5 ของประชากรที่เกิดก่อนจะเริ่มมีการฉีดวัคซีนป้องกันการติดเชื้อตับอักเสบ บี (ปี พ.ศ.2535) ส่วนประชากรที่เกิดหลัง พ.ศ.2535 ซึ่งได้บรรจุวัคซีนป้องกันไวรัสตับอักเสบชนิด บีในแผนการสร้างเสริมภูมิคุ้มกันโรคด้วยวัคซีนของประเทศ พบอัตราความชุกร้อยละ 0.6 ในภาพรวม ซึ่งจะมีจำนวนผู้ติดเชื้อไวรัสตับอักเสบ บีเรื้อรัง ประมาณ 2.2-3 ล้านคน

จากที่กล่าวมาข้างต้น ผู้วิจัยได้ตระหนักถึงปัจจัยต่าง ๆ ที่มีผลต่อการเกิดการติดเชื้อไวรัสตับอักเสบ เพื่อลดขั้นตอนและลดระยะเวลาในการปฏิบัติงาน จึงได้นำกระบวนการเหมืองข้อมูล (Data Mining) มาช่วยในการพิจารณาการเกิดการติดเชื้อไวรัสตับอักเสบ

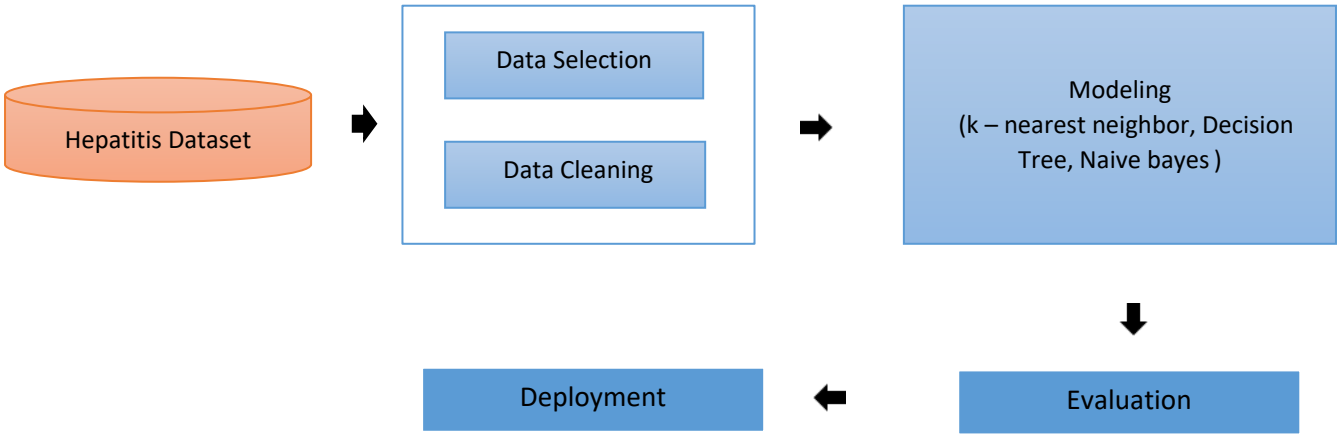
2. วัตถุประสงค์การวิจัย

2.1 เพื่อศึกษาแนวทางในการแนะนำการป้องกันการเกิดไวรัสตับอักเสบ

2.2 เพื่อเปรียบเทียบประสิทธิภาพในการจำแนกโรคไวรัสตับอักเสบด้วยเทคนิค KNN , Decision Tree , Naïve bayes

3. วิธีดำเนินการวิจัย

3.1 กรอบแนวคิดในการวิจัย แสดงเป็นรูปภาพขั้นตอนการออกแบบการวิจัย



ภาพที่ 1 กรอบแนวคิดในการวิจัย

3.2 ข้อมูลสำหรับการวิจัย

ข้อมูลสำหรับการทำวิจัยเป็นข้อมูลที่นำมาจาก UCI โดยมีชื่อไฟล์ว่า hepatitis เป็นข้อมูลเกี่ยวกับชุดข้อมูลจำแนกการมีชีวิตกับไม่มีชีวิตจากการเป็นไวรัสตับอักเสบ โดยมีข้อมูลภายใน 19 คอลัมน์ 113 แถว คอลัมน์ที่ 1 ชื่อว่า Class เป็นการจำแนกการมีชีวิตหรือไม่มีชีวิต คอลัมน์ที่ 2 ชื่อว่า Age เป็นคุณลักษณะอายุ คอลัมน์ที่ 3 ชื่อว่า Sex เป็นคุณลักษณะเพศ คอลัมน์ที่ 4 ชื่อว่า Steroid เป็นคุณลักษณะ Steroid คอลัมน์ที่ 5 ชื่อว่า Antivirals เป็นคุณลักษณะ Antivirals คอลัมน์ที่ 6 ชื่อว่า Fatigue เป็นคุณลักษณะความเหนื่อยล้า คอลัมน์ที่ 7 ชื่อว่า Malaise เป็นคุณลักษณะความไม่สบาย คอลัมน์ที่ 8 ชื่อว่า Anorexia เป็นคุณลักษณะอาการเบื่ออาหาร คอลัมน์ที่ 9 ชื่อว่า Liver Big เป็นคุณลักษณะตับใหญ่ คอลัมน์ที่ 10 ชื่อว่า Liver Firm เป็นคุณลักษณะตับแข็ง คอลัมน์ที่ 11 ชื่อว่า Spleen Palpable เป็นคุณลักษณะความชัดเจนของม้าม คอลัมน์ที่ 12 ชื่อว่า Spiders เป็นคุณลักษณะแมงมุม คอลัมน์ที่ 13 ชื่อว่า Ascites เป็นคุณลักษณะน้ำในช่องท้อง คอลัมน์ที่ 14 ชื่อว่า Varices เป็นคุณลักษณะความแปรปรวน คอลัมน์ที่ 15 ชื่อว่า Bilirubin เป็นคุณลักษณะบิลิรูบิน คอลัมน์ที่ 16 ชื่อว่า Alk Phosphate เป็นคุณลักษณะอัลคาไลน์ ฟอสฟาเทสคอลัมน์ที่ 17 ชื่อว่า Sgot เป็นคุณลักษณะ Sgot คอลัมน์ที่ 18 ชื่อว่า Albumin เป็นคุณลักษณะแอลบูมิน คอลัมน์ที่ 19 ชื่อว่า Histology เป็นคุณลักษณะ Histology

3.3 ขั้นตอนในการดำเนินการวิจัย (ขั้นตอนการทำ Data mining-KDD)

3.3.1 สำรวจหาข้อมูลที่ต้องการ และเหมาะสมในการทำวิจัย

3.3.2 ทำการสำรวจและเก็บรวบรวมข้อมูลวิจัยที่เกี่ยวข้องกับการใช้เทคนิคการทำเหมืองข้อมูล เพื่อศึกษาแนวทางในการทำวิจัย

3.3.3 ศึกษาข้อมูลที่เลือกมาและทำการ Clean Data ข้อมูลที่ไม่เกี่ยวข้องออกไป โดยการลบ missing value และ ลบ คอลัมน์ PROTIME

3.3.4 จากนั้นนำข้อมูลที่ Clean Data แล้ว เข้า Google Collaboratory และใช้ Data Mining Techniques ที่เลือกมา คือ k – nearest neighbor, Decision Tree, Naive bayes เพื่อเปรียบเทียบประสิทธิภาพ

3.3.5 ทำการประเมินผล(Evaluation) เพื่อนำผลที่ได้จากการนำเข้า Google Collaboratory มาพิจารณาว่าเทคนิคเหมืองข้อมูลชนิดไหนมีประสิทธิภาพที่เหมาะสมที่สุด

3.3.6 ทำการปรับใช้ (Deployment) เพื่อให้งานวิจัยเกิดประสิทธิภาพมากขึ้น และอาจจะนำผลที่ได้มาใช้ให้เกิดประโยชน์เกี่ยวกับการให้คำแนะนำเรื่องโรคไวรัสตับ

3.4 การประเมินผลการวิจัย

การวิจัยนี้ใช้วิธีการประเมินประสิทธิภาพ ได้แก่ Accuracy คือ มาตรวัดค่าความแม่นตรง คือ ค่าที่บอกว่าการทำนาย ทำนายได้แม่นยำแค่ไหน สามารถคำนวณได้ดังนี้

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

โดยที่ TP คือค่า True Positive, TN คือค่า True Negative,
FP คือค่า False Positive, FN คือค่า False Negative

4. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

4.1 ทฤษฎีที่เกี่ยวข้อง

4.1.1 การทำเหมืองข้อมูล (Data mining) เป็นกระบวนการในการสกัดเพื่อค้นหารูปแบบหรือความรู้ จากฐานข้อมูลขนาดใหญ่ เพื่อให้ได้สารสนเทศที่น่าสนใจที่ไม่คาดคิดมาก่อนและเป็นประโยชน์มาใช้ สนับสนุนการตัดสินใจในองค์กร

4.1.2 เพื่อนบ้านใกล้ที่สุดเค (k – nearest neighbor) หรือ เคเอ็นเอ็น (KNN) เป็นกระบวนการเรียนรู้แบบมีผู้สอนสำหรับปัญหาการจำแนกประเภทข้อมูลที่ถือว่าเป็นวิธีการพื้นฐานและง่ายที่สุดวิธีการหนึ่งในศาสตร์การเรียนรู้ของเครื่อง โดยจัดเป็นการเรียนรู้แบบขี้เกียจ (lazy learning) กล่าวคือ ไม่มีการสร้างตัวแบบจากชุดข้อมูลสอน แต่จะเริ่มคำนวณเมื่อมีการสอบถามเข้ามา

4.1.3 ต้นไม้ตัดสินใจ (decision tree) เป็นกระบวนการเรียนรู้แบบมีผู้สอนสำหรับปัญหาการจำแนกประเภทข้อมูลโดยใช้ทฤษฎีสารสนเทศ (information theory) และโครงสร้างต้นไม้ (tree structure) ในการสร้างตัวแบบ

4.1.4 Naive Bayes คือ การทำเหมืองข้อมูลในแบบ classifier ที่ถูกสร้างขึ้นโดยหลักความ น่าจะเป็น Naïve Bayesian Classificationจะใช้วิเคราะห์หาความน่าจะเป็นของสิ่งที่ยังไม่เคยเกิดขึ้น โดยการคาดเดาจากสิ่งที่เคยเกิด ขึ้นมาก่อน

4.1.5 ไวรัสตับอักเสบ (Hepatitis) คือ การที่ตับมีอาการอักเสบ และติดเชื้อ

4.2 งานวิจัยที่เกี่ยวข้อง

4.2.1 Pratheuang Wongthong et al., [1] การประยุกต์เทคนิคเหมืองข้อมูลค้นหาลักษณะนิสัย ของผู้ประกอบอาชีพด้านคอมพิวเตอร์ ได้ทำการศึกษาเพื่อสร้างและเปรียบเทียบประสิทธิภาพตัวแบบความสัมพันธ์ของข้อมูล จากเทคนิค Neural Networks, Naive Bayes และ Decision Tree เพื่อค้นหาลักษณะนิสัยของ ผู้ประกอบอาชีพด้านคอมพิวเตอร์ตามทฤษฎีการเลือกอาชีพของฮอลแลนด์โดยพิจารณาจากตัวแบบ จากค่าความถูกต้อง (Accuracy) และความแม่นยำ (Precision) ดำเนินการเก็บรวบรวมข้อมูลจากผู้ประกอบอาชีพด้านคอมพิวเตอร์ ใช้การสุ่มตัวอย่างจาก จำนวนผู้ที่สำเร็จการศึกษาระดับปริญญาตรี หลักสูตรวิศวกรรมคอมพิวเตอร์ วิทยาการคอมพิวเตอร์ และวิทยาศาสตร์คอมพิวเตอร์ กลุ่มตัวอย่าง 395 คน หลังจากรวบรวมข้อมูล แล้วจัดเตรียมข้อมูลให้ อยู่ในรูปแบบไฟล์ CSV จากนั้นวิเคราะห์ด้วยวิธี 5 Fold-Cross-Validation แล้วสร้างตัวแบบทั้ง 3 เทคนิค จากนั้นวัดค่าประสิทธิภาพและค่าความแม่นยำของตัวแบบ

4.2.2 Jiraporn Jareanying [2] การพยากรณผลสัมฤทธิ์ทางการเรียน ด้วยเทคนิคเหมืองข้อมูลโดยใช้Rapid Mine ได้ทำการศึกษาและนำเสนอการทำเหมืองข้อมูลทางการศึกษา โดยการจำแนกประเภทข้อมูลการวิเคราะห์องค์ประกอบหลัก

ของข้อมูล เพื่อหาความสัมพันธ์ของตัวแปร และเปรียบเทียบประสิทธิภาพของอัลกอริทึม ซึ่งเป็นการศึกษาด้วยเทคนิค ต้นไม้ตัดสินใจ เทคนิคป่าแห่งการทำนาย การเรียนรู้เบย์ และ K-NN โดยใช้ข้อมูลของนักเรียนระดับมัธยมศึกษาในโรงเรียนโปรตุเกส ประกอบด้วยข้อมูลด้านผลการเรียน ด้านความเป็นอยู่ และความเชื่อมโยงทางสังคมและโรงเรียนจาก UCI Machine Learning Repository มีข้อมูล 649 รายการ 31 แอตทริบิวต์

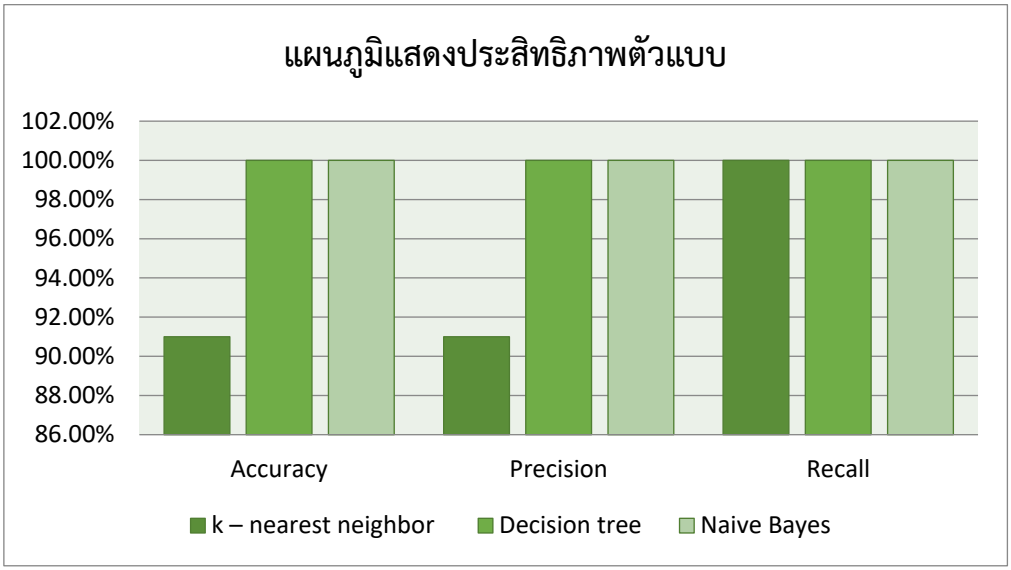
4.2.3 Rujiroj Hirose et al., [3] การวิเคราะห์พฤติกรรมที่มีผลต่อการซื้อผลิตภัณฑ์น้ำมันมะพร้าวในจังหวัดประจวบคีรีขันธ์ โดยใช้เทคนิคเหมืองข้อมูล โดยมีวัตถุประสงค์เพื่อ 1). ศึกษาและวิเคราะห์ปัจจัยที่ส่งผลต่อพฤติกรรมการซื้อผลิตภัณฑ์น้ำมันมะพร้าวในจังหวัดประจวบคีรีขันธ์ 2). สร้างโมเดลทำนายโอกาสการซื้อผลิตภัณฑ์น้ำมันมะพร้าวในจังหวัดประจวบคีรีขันธ์ โดยใช้เทคนิคต้นไม้การตัดสินใจ 3). สร้างกฎความสัมพันธ์ในการวิเคราะห์ความต้องการซื้อผลิตภัณฑ์ของลูกค้าผลการวิจัยพบว่า ปัจจัยที่ส่งผลต่อพฤติกรรมการซื้อสินค้าของลูกค้ามากที่สุด 3 อันดับแรก ได้แก่ ราคา ระดับการศึกษา และรายได้ ส่วนของโมเดลทำนายโอกาสการซื้อผลิตภัณฑ์น้ำมันมะพร้าวในจังหวัดประจวบคีรีขันธ์ โดยใช้เทคนิคต้นไม้การตัดสินใจ ผ่านวิธีการแบ่งกลุ่มข้อมูลแบบ Cross Validationมีค่าความถูกต้องมากที่สุด เท่ากับ 69.50 % และผลการวิเคราะห์ความต้องการซื้อผลิตภัณฑ์ของลูกค้า พบว่า ผลิตภัณฑ์ที่ลูกค้ามีความต้องการซื้อสูงสุด คือ สบู่น้ำมันมะพร้าว โดยมีค่า Support =31.2% และสินค้าที่ลูกค้ามักซื้อพร้อมกันมากที่สุดคือ โลชั่นทาตัวน้ำมันมะพร้าวและน้ำหอมน้ำมันมะพร้าวจากการโดยมีค่า support =2.7% และค่า Confidentเท่ากับ 64%

4.2.4 Supatsara Somjetna and Jaree Thongkham [4] การเปรียบเทียบประสิทธิภาพเทคนิคเหมืองข้อมูลในการสร้างแบบจำลองเพื่อจำแนกความคิดเห็น ของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร งานวิจัยฉบับนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของเทคนิคในเหมืองข้อมูลในการสร้างแบบจำลองจำแนกความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร งานวิจัยนี้ใช้6 เทคนิค ได้แก่ เทคนิคริปเปอร์เทคนิคต้นไม้ตัดสินใจแบบซี4.5 เทคนิคนาอีฟเบย์ เทคนิคซัพพอร์ตเวคเตอร์แมชชีนเทคนิคเคเนียร์เรสเนเบอร์และเทคนิคต้นไม้ป่าสุ่มมาสร้างแบบจำลองความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร โดยข้อมูลนั้นถูกรวบรวมมาเฉพาะความคิดเห็นของผู้ปกครองที่มีลักษณะเป็นข้อความภาษาไทยบนเครือข่ายสังคมออนไลน์ผ่านเว็บไซต์พันทิปและเฟซบุ๊คจำนวนทั้งหมด 1, 925 ข้อความ ตัวแปรที่ใช้ในงานวิจัยนี้ผู้วิจัยได้เลือกเฉพาะคำวิเศษณ์ที่สามารถระบุความรู้สึกได้เป็นอย่างดี 10-โฟลด์ครอสวาลิเดชั่นได้ถูกนำมาใช้ในการแบ่งกลุ่มข้อมูลชุดเรียนรู้และชุดทดสอบ รวมถึงวัดประสิทธิภาพของแบบจำลองด้วยโดยใช้ค่าความแม่นยำค่าความระลึก และค่าความถ่วงดุล หลังจากทำการทดลองและวัดประสิทธิภาพของแบบจำลอง พบว่าเทคนิคป่าสุ่มเป็นเทคนิคที่ดีที่สุดในการวิเคราะห์ความคิดเห็นสำหรับข้อมูลชุดนี้ โดยให้ค่าความแม่นยำ89.62% ค่าความระลึก78.38% และค่าความถ่วงดุล83.55%

4.2.5 Puth Muangpaisarn et al., [5] วิจัยความชุกและปัจจัยทำนายการเกิดพังผืดในตับที่มีนัยสำคัญประเมินโดย Transient Elastography ในผู้ป่วยไวรัสตับอักเสบบีเรื้อรังอายุ 40 ปีขึ้นไป มีวัตถุประสงค์เพื่อศึกษาความชุกและปัจจัยทำนายการเกิดพังผืดในตับที่มีนัยสำคัญซึ่งประเมินโดย transient elastography ในผู้ป่วยไวรัสตับอักเสบบีเรื้อรังที่มีอายุตั้งแต่ 40 ปีขึ้นไป ผลการวิจัยพบว่า 47.3% ของผู้ป่วยโรคไวรัสตับอักเสบบีเรื้อรังอายุ 40 ปีขึ้นไปที่ได้รับการประเมินโดย transient elastography มีความชุกของการเกิดพังผืดในตับอย่างมีนัยสำคัญ พังผืดในตับที่มีนัยสำคัญมีความสัมพันธ์อย่างมีนัยสำคัญทางสถิติกับเพศ, HBeAg ในเชิงบวก, anti-HBe ในเชิงลบ, HBV-DNA ในซีรั่ม≥20,000 IU/mL และกลุ่มโรคที่ใช้งานอยู่ เมื่อเปรียบเทียบกับพังผืดในตับที่ไม่มีนัยสำคัญ กลุ่มที่มีพังผืดในตับที่มีนัยสำคัญจะมีค่าเฉลี่ยของระดับเกล็ดเลือดและอัลบูมินต่ำกว่าอย่างมีนัยสำคัญทางสถิติ และค่าเฉลี่ยของ INR, Globulin, Bilirubin, AST, ALT และ ALP สูงกว่า รวมทั้งมีค่ามัธยฐานสูงกว่า ของซีรั่ม HBV-DNA นอกจากนี้ การวิเคราะห์การถดถอยโลจิสติกหลายตัวแปรแสดงเกล็ดเลือด <150,000/มม.^{3} (OR 7.35; 95%CI 1.64-33.0; p=0.01) และ INR>1.1 (OR 5.77; 95%CI 1.53-21.78; p =0.01) เป็นปัจจัยทำนายการเกิดพังผืดในตับที่มีนัยสำคัญ

5. ผลการวิจัย

ผลการจำแนกประเภทข้อมูลจากชุดข้อมูลการเกิดโรคไวรัสตับอักเสบผ่านการจำแนกข่อมูลด้วยเทคนิคเหมืองข้อมูล 3 วิธี ได้แก่ได้แก่ เพื่อนบ้านที่ใกล้ที่สุด ต้นไม้ตัดสินใจ และนาอีฟเบย์ โดยแบ่งข่อมูลในการเรียนรู้ (Training Data) และข่อมูลทดสอบ (Testing Data) ตามซอฟต์แวร์ Google Collaboratory ได้แก่ ค่าความแม่นยำ (Accuracy) ค่าความถูกต้อง (precision) และค่าความระลึก (recall) แสดงผลในรูปแบบตารางที่ 1 และภาพที่ 2



ภาพที่ 2 การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลการเกิดโรคไวรัสตับอักเสบ

ตารางที่ 1 เปรียบเทียบแสดงค่าความแม่นยำ ความถูกต้อง และค่าความระลึกในการจำแนกประเภทจากเทคนิคกับชุดข้อมูลการแปลงข้อมูล

เทคนิค	ความแม่นยำ	ความถูกต้อง	ความระลึก
k – nearest neighbor	91.00%	91.00%	100%
Decision tree	100%	100%	100%
Naive Bayes	100%	100%	100%

จากตารางที่ 1 และภาพที่ 2 การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลการเกิดโรคไวรัสตับอักเสบด้วยเทคนิค 3 วิธี ได้แก่ เพื่อนบ้านที่ใกล้ที่สุด ต้นไม้ตัดสินใจ และนาอีฟเบย์ ผลการทดลองทำให้ทราบว่าวิธีนาอีฟเบย์และต้นไม้ตัดสินใจให้ประสิทธิภาพที่ดีที่สุด โดยให้ค่าความแม่นยำ 100% ค่าความถูกต้อง 100% และค่าความระลึก 100% ส่วนรองลงมาคือ วิธีเพื่อนบ้านที่ใกล้ที่สุดให้ค่าความแม่นยำ 91% ค่าความถูกต้อง 91% และค่าความระลึก 100%

6. สรุปผล

ผลการวิจัยนี้ได้ทำการจำแนกข้อมูลด้วยวิธีการใช้เทคนิคเหมืองข้อมูล 3 วิธีการ ได้แก่ 3 เทคนิควิธี เพื่อนบ้านที่ใกล้ที่สุด ต้นไม้ตัดสินใจ และนาอีฟเบย์ จากชุดข้อมูลโรคไวรัสตับอักเสบ ซึ่งผลการประเมินประสิทธิภาพตัวแบบ คือ ผลการทดลองทำให้ทราบว่าวิธีนาอีฟเบย์และต้นไม้ตัดสินใจให้ประสิทธิภาพที่ดีที่สุด โดยให้ค่าความแม่นยำ 100% ค่าความถูกต้อง 100% และค่าความระลึก 100% ส่วนรองลงมาคือ วิธีเพื่อนบ้านที่ใกล้ที่สุดให้ค่าความแม่นยำ 91% ค่าความถูกต้อง 91%

และค่าความระลึก 100% จึงสรุปได้ว่า วิธีต้นไม้ตัดสินใจ และนาอีฟเบย์เป็นเทคนิคที่เหมาะสมที่สุดที่จะนำมาใช้จำแนกข้อมูลการโรคไวรัสตับอักเสบชุดนี้

7. เอกสารอ้างอิง

[1] Pratheuang Wongthong, Varit Kankaew, Aphichaya Khwankaew, and Yuvadee Chomdaeng (2020). Applied of Data Mining Techniques for Searching Characteristics of Computer Career. [Online] Retrieved on 27th February 2023 from https://research.rmutsb.ac.th/fullpaper/2562/research.rmutsb-2562-20201006100035839.pdf .(In Thai)

[2] Jiraporn Jareanying (2020). The Prediction of Student Performance using Data Mining Techniques with Rapid Miner. [Online] Retrieved on 27th February 2023 from http://irithesis.swu.ac.th/dspace/bitstream/123456789/1231/1/gs591130025.pdf .(In Thai)

[3] Rujiroj Hirose, Peangruethai Nusawat, Kannika Boonkasem, and Noppadol Saiktikorn (2021). Behavior analysis affecting purchasing coconut oil products, in Prachuap Khiri Khan Province using data. Rattanakosin Journal of Science and Technology, 2(3), 26–46. Retrieved from https://ph02.tci-thaijo.org/index.php/RJST/article/view/243476 .(In Thai)

[4] Supatsara Somjetna and Jaree Thongkham (2020). Performance Comparison of Data Mining Techniques for Building Classification Models of the Parent Toward Children who use Smart Phone. [Online] Retrieved on 27th February 2023 from https://li01.tci thaijo.org/index.php/sci_ubu/article/view/247842. (In Thai)

[5] Puth Muangpaisarn, Theerapat Orprayoon, and Jaruwan Chanyaswad. (2021). Prevalence and Predicting Factors of Significant Liver Fibrosis Evaluated by Transient Elastography in Patients with Chronic Hepatitis B Aged 40 Years or More. [Online] Retrieved on 27th February 2023 from https://he02.tci-thaijo.org/index.php/ppkjournal/article/view/251403 .(In Thai)