การเพิ่มประสิทธิภาพจำแนกข้อมูลผลกระทบโควิด-19 ต่อผู้ป่วยมะเร็งตับ

Optimization Classification of Covid-19 effect on Liver Cancer

ธรรมนูญ ปัญญาทิพย์[1*], ปนัดดา โพธินาม[2] และ คมกริช อ่อนประสงค์[3]

Tammanoon Panyatip[1*], Panatda Phothinam[2] and Komgrich Onprasonk[3]

สาขาวิชาวิทยาการสารสนเทศและคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยีสุขภาพ มหาวิทยาลัยกาฬสินธุ์[1*,2,3]

E-Mail : tammanoon@ksu.ac.th[1*], panatdapo@hotmail.com[2], onprasonk@gmail.com[3]

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อ 1) จำแนกข้อมูลผลกระทบโควิด-19 ต่อผู้ป่วยมะเร็งตับ และ 2) เพิ่มประสิทธิภาพการจำแนกข้อมูลด้วยการเรียนรู้แบบรวมกลุ่มข้อมูลผลกระทบโควิด-19 ต่อผู้ป่วยมะเร็งตับ ด้วยเทคนิคการเรียนรู้ของเครื่อง โดยมีวิธีการจำแนกข้อมูลด้วยอัลกอริทึม ประกอบด้วย 1) ซัพพอร์ตเวกเตอร์แมทชีน 2) แบบนาอีฟเบย์ 3) เพื่อนบ้านใกล้ที่สุด 4) ต้นไม้ตัดสินใจ 5) โครงข่ายประสาทเทียม และการเพิ่มประสิทธิภาพการจำแนกข้อมูล ประกอบด้วย 1) การโหวต 2) การสุ่มป่าไม้ ชุดข้อมูลทดลองวิจัย ได้แก่ ข้อมูลโควิด-19 ผลกระทบต่อผู้ป่วยมะเร็งตับ แบ่งชุดข้อมูลการทดลองเป็น 2 วิธี คือ แบบแยกชุดข้อมูลเป็น 80:20 และแบ่งแบบ K-Fold Cross Validation เครื่องมือที่ใช้ในการทดลองภาษาไพทอนและโปรแกรม Visual Studio Code วัดประสิทธิภาพด้วยวิธี ค่าความแม่นยำ ค่าความครบถ้วน F1-Score และค่าความถูกต้อง

ผลการวิจัย พบว่า 1) การจำแนกข้อมูลผลกระทบโควิด-19 ต่อผู้ป่วยมะเร็งตับ ด้วยการแบ่งชุดข้อมูล 80:20 วิธีการจำแนกที่ดีที่สุด ได้แก่ วิธีการต้นไม้ตัดสินใจ และวิธีการโครงข่ายประสาทเทียม มีความถูกต้อง 100% การแบ่งชุดข้อมูลทดสอบประสิทธิภาพ 5-Fold Cross Validation วิธีการจำแนกที่ดีที่สุด ได้แก่ วิธีการต้นไม้ตัดสินใจ มีความถูกต้อง 99.6% และ 2) ผลการเพิ่มประสิทธิภาพการจำแนกข้อมูลด้วยการเรียนรู้แบบรวมกลุ่มทดสอบประสิทธิภาพ 5-Fold Cross Validation วิธีการสุ่มป่าไม้ มีความถูกต้อง 99.8%

คำสำคัญ : จำแนกข้อมูล, โควิด19, มะเร็งตับ, การเรียนรู้ของเครื่อง

ABSTRACT

This research aims to 1) identify COVID-19 impact data and 2) enhance data classification by learning to integrate COVID-19 impact data. The method of data classification with algorithms includes 1) Support Vector Machines, 2) Naïve Bayes, 3) K-Nearest Neighbor, 4) Decision Tree 5) Artificial Neural Networks. Data classification optimization consists of 1) Ensemble Vote and 2) Random Forest. Research datasets consisted of COVID-19 that affect liver cancer patients by dividing the experimental dataset into a separate 80:20 dataset and a K-Fold Cross Validation. Tools used in Python language experiments and visual studio code. Performance is measured using Precision, Recall, F1-Score, and accuracy.

The results showed that: 1) the classification of COVID-19 effect on liver cancer, dividing the 80:20 dataset of the best classification methods including decision tree and neural network, the accuracy of 100%, 5-fold cross-validation performance test dataset division. The best classification includes the decision tree, accuracy of 99.6%. 2) Data classification optimization with the ensemble, 5-fold cross-validation, random forest. It has an accuracy of 99.8%.

Keywords : Classification, Covid-19, Liver Cancer, Machine Learning

บทนำ

การแพร่ระบาดของโรคโควิด-19 (COVID-19) นับตั้งแต่เริ่มต้นของการระบาดใหญ่ของ โควิด-19 มีผู้ติดเชื้อที่ยืนยันแล้วกว่า 629 ล้านคน และมีผู้เสียชีวิต 6.59 ล้านคนทั่วโลก จากปัญหาการแพร่ระบาดของโควิด-19 ส่งผลให้บริการด้านสุขภาพสำหรับผู้ป่วยโรคเรื้อรัง เช่น โรคมะเร็ง โควิด-19 ยังทำให้เกิดความผิดปกติของอวัยวะ การรักษามะเร็งที่ซับซ้อน ในประเทศส่วนใหญ่ที่มีการระบาดของโควิด-19 มีการปรับเปลี่ยนการจัดการมะเร็งได้ถูกนำมาใช้เพื่อรองรับวิกฤตการณ์และลดความเสี่ยงของผู้ป่วยมะเร็งต่อการติดเชื้อ ในประเทศที่ตัวเลขโควิด-19 กำลังลดลง ทีมแพทย์เตรียมพร้อมที่จะกลับมาปฏิบัติตามปกติอย่างค่อยเป็นค่อยไป [1] ซึ่งก่อนหน้านี้กิจกรรมด้านการดูแลสุขภาพหลายอย่าง เช่น การจัดการโรคเรื้อรัง การตรวจคัดกรองมะเร็ง และการรักษามะเร็งถูกยกเลิกหรือล่าช้า ด้วยเหตุนี้ การเสียชีวิตที่เกี่ยวข้องกับมะเร็งจึงเพิ่มขึ้น

ผลกระทบทั้งหมดจากการระบาดของโควิด-19 ต่อผู้ป่วยมะเร็งตับ (hepatocellular carcinoma : HCC) จากข้อมูลปัจจุบัน 10-40% ของผู้ป่วยที่ติดเชื้อโควิด-19 มีอาการบาดเจ็บที่ตับ กลไกหลายอย่างมีส่วนทำให้เกิดการบาดเจ็บที่ตับ รวมถึงการที่ไวรัสเข้าสู่เซลล์ตับ/ถุงน้ำดีโดยตรง ตับอักเสบจากภูมิคุ้มกัน การขาดออกซิเจน และความเป็นพิษต่อตับจากยาโดยตรง ในผู้ป่วย HCC โควิด-19 อาจทำให้โรคตับเรื้อรังที่มีอยู่รุนแรงขึ้นและทำให้การจัดการมะเร็งยุ่งยากขึ้น ผู้ป่วยมะเร็งมักมีความเสี่ยงในการติดเชื้อและผลลัพธ์ที่แย่ลง โดยเฉพาะผู้ที่เพิ่งได้รับการรักษามะเร็ง การตัดสินใจในการรักษา HCC ควรสมดุลกับการพิจารณาความพร้อมของทรัพยากรทางการแพทย์ ระดับความเสี่ยงในการติดเชื้อของโควิด-19 และอัตราส่วนความเสี่ยงและผลประโยชน์ของผู้ป่วยแต่ละราย ในพื้นที่การระบาดของโควิด-19 คลี่คลายลง แพทย์ควรเตรียมพร้อมที่จะจัดการกับผู้ป่วย HCC ที่มีภาระโรคและภาวะแทรกซ้อนที่สูงขึ้น [1] และทางตอนเหนือของอังกฤษเป็นตัวอย่างหนึ่ง ที่ประสบปัญหาจำนวนผู้ป่วยมะเร็งระยะแรกเพิ่มขึ้นอย่างต่อเนื่อง ส่วนใหญ่รับผลกระทบจากการระบาดของโควิด-19 [2]

ข้อมูลผลกระทบโควิด-19 ต่อผู้ป่วยมะเร็งตับ [3] ถูกรวบรวมจากทีมสหสาขาวิชาชีพตับและตับอ่อนของ Newcastle-upon-Tyne NHS Foundation Trust (NUTH) ในช่วง 12 เดือนแรกของการระบาดใหญ่ (มีนาคม 2020-กุมภาพันธ์ 2021) เปรียบเทียบกับการสังเกตย้อนหลัง กลุ่มผู้ป่วยต่อเนื่องในช่วง 12 เดือนก่อนหน้านั้น (มีนาคม 2019-กุมภาพันธ์ 2020) รวมผู้ป่วยรายใหม่ทั้งหมดที่มีการวินิจฉัยโรคมะเร็งตับ (HCC) หรือมะเร็งท่อน้ำดีภายในตับ (intrahepatic cholangiocarcinoma : ICC) ที่ได้รับการยืนยันทางรังสีวิทยาหรือการตรวจชิ้นเนื้อตามแนวทางสากล

การจำแนกข้อมูล (Classification) เป็นการจัดการประเภทข้อมูลแยกตามลักษณะของข้อมูลด้วยเทคนิคการเรียนรู้ของเครื่อง ซึ่งปัจจุบันมีเทคนิคการเรียนรู้ของเครื่องที่นำมาใช้ในการจำแนกข้อมูลมีหลายวิธี เช่น ซับพอร์ตเวคเตอร์แมทชีน (Support Vector Machines) นาอีฟเบย์ (Naïve Bayes) เพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbor) ต้นไม้ตัดสินใจ (Decision Tree) และโครงข่ายประสาทเทียม (Artificial Neural Networks) เป็นต้น แต่การจำแนกข้อมูลในบางครั้งยังไม่มีประสิทธิภาพหรือมีประสิทธิภาพต่ำ ทั้งนี้ขึ้นกับชุดข้อมูล (Dataset) ลักษณะข้อมูล การเตรียมข้อมูล การทำความสะอาดข้อมูล และการเลือกวิธีการจำแนกข้อมูลให้เหมาะสม

ดังนั้น ผู้วิจัยจึงได้ทำการทดลองข้อมูลด้วยเทคนิคการเรียนรู้ของเครื่อง การจำแนกข้อมูลจากข้อมูลผลกระทบโควิด-19 ต่อผู้ป่วยมะเร็งตับ ซึ่งจะช่วยพยากรณ์ความถูกต้องการจำแนกข้อมูลผู้ป่วยมะเร็งตับจากผลกระทบโควิด-19 และหาวิธีการเพิ่มประสิทธิภาพการจำแนกข้อมูลให้มีประสิทธิภาพมากยิ่งขึ้น

1. วัตถุประสงค์การวิจัย

1.1 เพื่อจำแนกข้อมูลผลกระทบโควิด-19 ต่อผู้ป่วยมะเร็งตับ

1.2 เพื่อเพิ่มประสิทธิภาพการจำแนกข้อมูลด้วยการเรียนรู้แบบรวมกลุ่มข้อมูลผลกระทบโควิด-19 ต่อผู้ป่วยมะเร็งตับ

2. เอกสารและงานวิจัยที่เกี่ยวข้อง

Ghosh, Dasgupta และ Swetapadma [4] ได้ศึกษาเกี่ยวกับซัพพอร์ตเวคเตอร์แมทชีน (Support Vector Machines : SVM) ที่เป็นการจำแนกแบบเชิงเส้นและไม่เป็นเชิงเส้น เพื่อทำนายภัยพิบัติแผ่นดินไหวและสึนามิ พบว่ามีความถูกต้อง 77% Mohan และคนอื่น [5] ได้ศึกษา การปรับปรุงความแม่นยำ SVM ด้วยการจำแนกประเภท ชุดข้อมูลเป็น Iris เปรียบเทียบเคอร์เนลฟังชัน ด้วยวิธีการ Linear Kernel, Polynomial Kernel, Radial Basis Function (RBF) และ Sigmoid พบว่า RBF มีประสิทธิภาพมากที่สุด มีความถูกต้อง 97.3 และสามารถเรียงลำดับความถูกต้องจากมากไปน้อย ดังนี้ รองลงเป็น Linear Kernel มีความถูกต้อง 96.66 Polynomial Kernel มีความถูกต้อง 95.33 และ Sigmoid มีความถูกต้อง 88.66 ตามลำดับ

Harahap และคนอื่น ๆ [6] ได้ศึกษาการดำเนินการจำแนกนาอีฟเบย์ (Naïve Bayes) วิธีการทำนายการซื้อ ชุดข้อมูลเป็นรถยนต์ จำนวน 20 เรคคอร์ด ที่คาดการณ์การซื้อรถยนต์ พบว่า มีความถูกต้อง 75% Singh และคนอื่น [7] ได้ศึกษาการเปรียบเทียบระหว่าง Multinomial และ Bernoulli สำหรับจำแนกข้อความด้วยนาอีฟเบย์ ชุดข้อมูลบทความข่าว จำนวน 312 เรคคอร์ด มีการเตรียมข้อมูลการด้วย เปลี่ยนตัวพิมพ์ใหญ่ให้เป็นตัวพิมพ์เล็ก การทำ Tokenization การลบเครื่องหมายวรรคตอน และ Stop words พบว่า Multinomial มีประสิทธิภาพมากว่า Bernoulli มีความถูกต้อง 73.4% และ 69.15% ตามลำดับ Abbas และคนอื่น [8] ได้ศึกษาการจำแนกข้อความด้วย Multinomial นาอีฟเบย์ ชุดข้อมูลการรีวิวภาพยนต์ ดำเนินการโดยใช้สถิติ การประมวลภาษาธรรมชาติ และการเรียนรู้ของเครื่อง สำหรับการแยกหมวดหมู่เนื้อหาของข้อความ มีการเตรียมข้อมูลด้วยวิธีการ ทำความสะอาดข้อมูล การทำ Word Tokenization, Bag of Word และ TF-IDF พบว่า มีความถูกต้อง 93.1%

Xing และ Bei [9] ได้ศึกษาการจำแนกข้อมูลขนาดใหญ่ด้านสุขภาพการแพทย์ด้วยอัลกอริทึมเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbor : KNN) เมื่อกำหนด K=9 อัลกอริทึม KKN มีความถูกต้อง 78% ผู้วิจัยได้มีการปรับปรุงอัลกอริทึม KKN เพื่อเพิ่มประสิทธิภาพ ด้วยการคำนวณค่าระยะทางด้วยวิธีการ European distance และแบบโคไซน์ (cosine) จัดการความหนาแน่นของข้อมูลด้วยการครอบตัดลดความหนาแน่นของคลัสเตอร์ หลังการปรับปรุงมีความถูกต้อง 85%

Patel และ Prajapati [10] ได้ศึกษาการจำแนกข้อมูลด้วยต้นไม้ตัดสินใจ ชุดข้อมูลการทำลองเป็นชุดข้อมูลรถยนต์ ประกอบด้วย จำนวน 1728 ชุดข้อมูล และ 6 แอตทริบิวต์ ทำการเปรียบเทียบอัลกอริทึม ID3, C4.5 CART พบว่า อัลกอริทึม CART มีประสิทธิภาพมากที่สุด มีความถูกต้อง 97.11% รองลงมาเป็นอัลกอริทึม C4.5 มีความถูกต้อง 92.36% และอัลกอริทึมID3 มีความถูกต้อง 89.35%

Al-Massri และคนอื่น ๆ [11] ได้ศึกษาการทำนายการจำแนกมะเร็ง SBRCTs (Small Blue Round Cell Tumors) โดยใช้โครงข่ายประสาทเทียม (Artificial Neural Network : ANN) พบว่า ดีเอ็นเอ SBRCTs มะเร็งเป็นโรคที่อันตรายมากทั่วโลก โมเดลโครงข่ายประสาทเทียมเป็นระบบการวินิจฉัยที่บรรลุระดับความแม่นยำการพยากรณ์โรคมะเร็ง ดีเอ็นเอ SBRCTs ซึ่งสามารถช่วยแพทย์ในการวางแผนที่จะปรับปรุงยา และให้การวินิจฉัยแก่ผู้ป่วยได้ทันท่วงที โมเดล ANN มีประสิทธิภาพความถูกต้อง 100%

Mohana และคนอื่น ๆ [12] ได้ศึกษาอัลกอริทึมการสุ่มป่าไม้ (Random Forest) สำหรับการจำแนกตามกลุ่มต้นไม้ (tree based ensemble) ด้วยวิธีการ Classification by ensembles from Random partitions (CERP) ข้อมูลเป็นชุดข้อมูลมะเร็งต่อมลูกหมาก ผ่านกระบวนการเรียนรู้ของเครื่อง ด้วยกระบวนการตรวจสอบข้อมูลและทำความสะอาดข้อมูล ผลการศึกษาพบว่า การจำแนกข้อมูลด้วยวิธีการ CERP มีความถูกต้อง 0.8333 และทำงานได้ดีสำหรับข้อมูลมิติสูง ซึ่งเป็นทางเลือกที่น่าสนใจเพราะไม่มี Bias ในการเลือกเมื่อสร้างต้นไม้ตัดสินใจ

วิธีดำเนินการวิจัย

1. เครื่องมือการวิจัย

1.1 เครื่องมือที่ใช้ในการวิจัย ได้แก่ ภาษาไพทอน และโปรแกรม Visual Studio Code

1.2 ชุดข้อมูลการทดลองวิจัย (Dataset) ได้แก่ COVID-19 effect on Liver Cancer [3] จากเว็บไซต์ https://www.kaggle.com/ ข้อมูล ณ วันที่ 26 กันยายน 2565

2. กลุ่มเป้าหมาย

ข้อมูลโควิด-19 กระทบต่อผู้ป่วยมะเร็งตับ มีการแบ่งชุดข้อมูล เพื่อใช้ในการจำแนกข้อมูล ประกอบด้วย

2.1 แบ่งชุดข้อมูลออกเป็น 2 ส่วน คือชุดบการเรียนรู้ (Train) 80 เปอร์เซ็นต์ และชุดสำหรับการทดสอบ (Test) 20 เปอร์เซ็นต์

2.2 K-Fold Cross Validation โดยผู้วิจัยได้แบ่งชุดข้อมูลจำนวน 5 ชุดข้อมูล เพื่อใช้สำหรับสลับเป็นชุดเรียนรู้ และชุดการทดสอบ

3. ขั้นตอนการดำเนินการวิจัย

3.1 การเลือกข้อมูลและตรวจสอบข้อมูล

3.1.1 การเลือกข้อมูล เป็นข้อมูลโควิด19 ที่มีผลกระทบต่อผู้ป่วยมะเร็งตับ และลักษณะข้อมูล (Raw Data) มีจำนวน 27 คอลัมน์ และมีข้อมูลจำนวน 450 แถว

	Cancer	Year	Bleed	Mode_Presentation	Gender	Etiology	Cirrhosis	HCC_TNM_Stage	HCC_BCLC_Stage	ICC_TNM_Stage	Treatment_grps	Alive_Dead
0	Y	Prepandemic	N	Surveillance	M	NAFLD	Y	II	A	NaN	Ablation	Alive
1	Y	Prepandemic	N	Surveillance	M	ARLD	Y	I	D	NaN	Supportive care	Dead
2	Y	Prepandemic	N	Surveillance	M	ARLD	Y	IV	B	NaN	Medical	Dead
3	Y	Prepandemic	N	Incidental	M	ARLD	Y	IV	C	NaN	Supportive care	Dead
4	Y	Prepandemic	N	Incidental	F	ARLD	Y	I	0	NaN	Supportive care	Alive

ภาพที่ 1 ลักษณะข้อมูล

3.2.2 ตรวจสอบข้อมูลที่ Outlier

3.3.3 ตรวจสอบความสัมพันธ์ของข้อมูล

3.2 การเตรียมข้อมูล

3.2.1 การทำความสะอาดข้อมูล (Clean Data) ประกอบด้วย

1) ลบคอลัมน์ที่ไม่สำคัญออก ได้แก่ 'Date_incident_surveillance_scan'

2) จัดการข้อมูลที่เป็นค่าว่าง (Nan)

3) ลบคอลัมน์ที่เป็นข้อมูล Label คือคอลัมน์ ‘Cancer’ ออกจากข้อมูลชุด Train

3.2.2 แปลงข้อมูลที่มีลักษณะตัวอักษร (Object) ให้เป็นตัวเลขทศนิยม (Float) ด้วยการเข้ารหัส วิธีการ Label Encoding และวิธีการ One Hot Encoding

	Year	Month	Bleed	Mode_Presentation	Age	Gender	Etiology	Cirrhosis	Size	HCC_TNM_Stage	...	Type_of_incidental_finding	Surveillance_programme
0	1.0	0.0	0.0	1.0	29.0	1.0	4.0	1.0	12.0	1.0	...	4.0	1.0
1	1.0	0.0	0.0	1.0	31.0	1.0	0.0	1.0	30.0	0.0	...	4.0	1.0
2	1.0	0.0	0.0	1.0	25.0	1.0	0.0	1.0	41.0	3.0	...	4.0	1.0
3	1.0	0.0	0.0	0.0	34.0	1.0	0.0	1.0	63.0	3.0	...	2.0	0.0
4	1.0	0.0	0.0	0.0	27.0	0.0	0.0	1.0	48.0	0.0	...	2.0	0.0

ภาพที่ 2 การแปลงข้อมูล

3.3 การจำแนกประเภทข้อมูล

3.3.1 ซัพพอร์ตเวกเตอร์แมทชีน (Support Vector Machines)

ซัพพอร์ตเวกเตอร์แมทชีน เป็นวิธีการจำแนกประเภทข้อมูลที่มีการระบุจุดกึ่งกลางไฮเปอร์เพลน (Hyper plane) ได้อย่างเหมาะสม เพื่อจำแนกข้อมูลเป็น 2 ส่วน SVM มีฟังก์ชันเคอร์เนล (Kernel) ให้เลือกใช้ในการจำแนกข้อมูล เช่น Linear Kernel, Polynomial Kernel, Radial Basis Function และ Sigmoid เป็นต้น ซึ่งต้องเลือกเคอร์เนลให้เหมาะสมกับการจำแนกข้อมูล โดยงานวิจัยนี้ใช้ Radial Basis Function (RBF) ดังงานวิจัยของ Mohan และคนอื่น [5]

3.3.2 แบบนาอีฟเบย์ (Naïve Bayes)

นาอีฟเบย์ [7] เป็นอัลกอริทึมจำแนกทางสถิติตามทฤษฎีของเบย์ (Bayes’ Theorem) ช่วยให้หาเงื่อนไขความน่าจะเป็นที่จะเกิดขึ้นของสองเหตุการณ์ขึ้นอยู่กับความน่าจะเป็นของเหตุการณ์แต่ละเหตุการณ์ อัลกอริทึมนี้ทำงานบนหลักการที่ว่าแต่ละแอตทริบิวต์เป็นอิสระจากตัวอื่น นาอีฟเบย์ ประกอบด้วย Gaussian สำหรับจัดการข้อมูลต่อเนื่อง, Multinomial แสดงถึงความถี่ของเหตุการณ์บางอย่างที่ถูกสร้างขึ้น และ Bernoulli สำหรับงานจำแนกข้อมูลเอกสารที่มีแอตทริบิวต์แบบไบนารี โดยงานวิจัยนี้ใช้ Multinomial ดังงานวิจัยของ Abbas และคนอื่น [8]

3.3.3 เพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbor)

อัลกอริทึมเพื่อนบ้านที่ใกล้ที่สุด (K-Nearest Neighbor : KNN) เป็นอัลกอริทึมในการแบ่งคลาสมีการคำนวณหาระยะทางที่ใกล้ที่สุด โดยปกติจะเป็นวิธีการวัดระยะทางแบบยูคลิเดียน (Euclidean distance) ซึ่งเป็นวัดระยะทางระหว่างจุด 2 จุดในแนวเส้นตรง หากกรณีข้อมูลที่ต้องการอยู่ในคลาสใด จะคำนวณระยะทางที่ใกล้มากที่สุดจะหาผลรวมจากการกำหนดค่า k โดยปกติจะกำหนดเป็นเลขคี่ ดังนั้น การหาค่า k สำหรับการจำแนกประเภท KNN ซึ่งเป็นค่าที่เหมาะสมกับข้อมูล เพื่อให้ได้อัลกอริทึม KNN มีประสิทธิภาพมากที่สุด [13] งานวิจัยนี้จึงได้มีการทดลองเพื่อหาค่า k ที่เหมาะสมที่สุด โดยการวัดประสิทธิภาพด้วยวิธี 5-fold cross validation ดังตารางที่ 1

ตารางที่ 1 เปรียบเทียบประสิทธิภาพ KNN

กำหนดค่า k	Precision	Recall	F1-score	Accuracy
3	0.785	0.816	0.800	0.718
5	0.754	0.848	0.797	0.704
7	0.771	0.858	0.812	**0.727**
9	0.744	0.861	0.798	0.700
11	0.743	0.903	0.815	0.718

จากตารางที่ 1 เปรียบเทียบประสิทธิภาพ KNN เพื่อหาค่า k ที่เหมาะสมมากที่สุด พบว่า k=7 มีความถูกต้องมากที่สุด 72.7% ดังนั้น งานวิจัยนี้จึงเลือกกำหนดค่า k มีค่าเท่ากับ 7

3.3.4 ต้นไม้ตัดสินใจ (Decision Tree)

ต้นไม้ตัดสินใจ [14] ใช้อัลกอริทึมที่หลากหลายในการตัดสินใจแบ่งโหนดหนึ่งโหนดออกเป็นสองโหนดย่อยขึ้นไปหรือโหนดลูก เช่นเดียวกับแผนผังการตัดสินใจแบ่งโหนดในตัวแปรที่มีอยู่ทั้งหมด แล้วเลือกการแยกซึ่งส่งผลให้ได้โหนดย่อยที่เหมือนกันมากที่สุด [15] อัลกอริทึมต้นไม้การตัดสินใจ ได้แก่ ID3, C4.5, CART ฯลฯ อัลกอริทึมต้นไม้การตัดสินใจที่เก่าที่สุด ID3 จะสร้างต้นไม้ตัดสินใจโดยเลือกแอตทริบิวต์ที่มีขนาดใหญ่ของข้อมูลมากที่สุด และค่าที่สำคัญสำหรับการแยกโหนด รองรับการประมวลผลข้อมูลแบบไม่ต่อเนื่อง และเทรนโมเดลมีแนวโน้มที่จะเกิด overfitting [16] อัลกอริทึม C4.5 เป็นการปรับปรุงอัลกอริทึม ID3 เพื่อป้องกันปรากฏการณ์ overfitting

ขั้นตอนจะทำการตัดแต่งกิ่งบนพื้นฐานของ ID3 กระบวนการดำเนินการคือการระบุเกณฑ์ (Threshold) เลือกตัวอย่างน้อยกว่าเกณฑ์ที่กำหนด ซึ่งช่วยลด overfitting ได้ ส่วน CART (Classification and Regression Trees) [17] คล้ายกับ C4.5 แต่แตกต่างกันตรงที่รองรับตัวแปรเป้าหมายที่เป็นตัวเลข (การถดถอย) และไม่คำนวณชุดกฎดำเนินการแบ่งส่วนออกสองทางเลือก ในข้อมูลชุดเทรนตามเกณฑ์ขั้นต่ำของ Gini แบ่งชุดตัวอย่างปัจจุบันออกเป็นชุดตัวอย่างย่อยสองชุด ดังนั้น แต่ละโหนดที่ไม่ใช่ลีฟของต้นไม้ตัดสินใจที่สร้างขึ้น มีสองโหนดเป็นต้นไม้ไบนารีและการจำแนกประเภทต้นไม้ตัดสินใจ [18] สรุปคือ CART เป็นต้นไม้ไบนารี ในขณะที่ ID3 และ C4.5 สามารถเป็นต้นไม้แบบหลายทางเลือกได้ ในงานวิจัยนี้ใช้ CART ดังงานวิจัยของ Patel และ Prajapati [10]

3.3.5 โครงข่ายประสาทเทียม (Artificial Neural Networks)

อัลกอริทึมที่จำลองการทำงานของโครงข่ายประสาทสมองของมนุษย์ ประกอบด้วย เซลล์ประสาท เรียกว่า นิวรอน เซลล์ประสาทในการรับกระแสประสาท เรียกว่า เดนไดร์ เพื่อส่งกระแสปลายประสาท เรียกว่า แอคซอน จากแนวคิดดังกล่าว ได้ปรับมาเป็นกระบวนการทำงานของโครงข่ายประสาทเทียม ซึ่งหมายถึงเครือข่ายของโหนดที่เชื่อมต่อถึงกัน [19] ประกอบด้วย input layer ในการรับข้อมูลมาที่ชั้นของ hidden layer จะเชื่อมต่อกับทุกโหนดใน input layer และ output layer โดยการเชื่อมต่อเรียกว่า เอจ (Edge) หลังจากโครงการประสาทเทียมมีการเรียนรู้ จะมีค่าน้ำหนัก (weights) การคำนวณค่าน้ำหนัก และฟังก์ชันการแปลงด้วยฟังก์ชันซิกมอยด์ (sigmoid function) ฟังก์ชันไฮเปอร์โบลิกแทนเจนต์ (hyperbolic tangent function)

3.4 โมเดลการเรียนรู้แบบรวมกลุ่ม

เทคนิคการเรียนรู้แบบรวมกลุ่ม (Ensemble model) เป็นการนำโมเดลการจำแนกข้อมูลหลาย ๆ โมเดลมาทำงานร่วมกัน เพื่อให้ได้ประสิทธิภาพที่ดีที่สุด ซึ่งเทคนิค Ensemble ประกอบด้วย Vote Ensemble, Bagging, Boosting, Random Forest สำหรับงานวิจัยนี้ใช้วิธี Vote Ensemble และวิธี Random Forest เท่านั้น

3.4.1 Vote Ensemble

เป็นการปรับปรุงประสิทธิภาพการจำแนกประเภทไบนารี สร้างโมเดลด้วยเทคนิคจำแนกข้อมูล อาทิเช่น ต้นไม้ตัดสินใจ ซัพพอร์ตเวคเตอร์แมชชีน เพื่อนบ้านใกล้ที่สุด โครงข่ายประสาทเทียม ใช้หลักการโหวตที่มีลักษณะคล้ายกันเป็นคำตอบ และจากงานวิจัยของ ABRO [20] เพิ่มประสิทธิภาพด้วย vote ensemble จากอัลกอริทึมนาอีฟเบย์ (Naive Baye) โครงข่ายประสาทเทียม (Artificial Neural Network) และ แบบจำลองลอจิสติกส์ (Logistic model tree) ซึ่งสามารถช่วยเพิ่มประสิทธิภาพความถูกต้องได้ งานวิจัยได้นำอัลกอริทึมประกอบด้วย 1) ซัพพอร์ตเวกเตอร์แมทชีน 2) แบบนาอีฟเบย์ 3) เพื่อนบ้านใกล้ที่สุด 4) ต้นไม้ตัดสินใจ 5) โครงข่ายประสาทเทียม ใช้ในการ vote ensemble สมการดังนี้

$$\hat{y} = \mathrm{mode}\{C_1(x), C_2(x), ..., C_m(x)\} \quad (1)$$

เมื่อ $\hat{y}$ เป็นการทำนายผ่านการลงคะแนนเสียงส่วนใหญ่ของแต่ละการจำแนกข้อมูลแต่ละชนิด C_m

3.4.2 การสุ่มป่าไม้ (Random Forest)

Random Forest เป็นอัลกอริทึมจำแนกประเภทการเรียนรู้ของเครื่อง พิจารณาการจำแนกประเภทจากโครงสร้างต้นไม้ ด้วยเวกเตอร์สุ่มอิสระที่แจกแจงเหมือนกัน และต้นไม้แต่ละต้นทำการโหวตหนึ่งหน่วยที่มีอินพุต x สำหรับคลาสที่นิยม [21] ซึ่ง Random Forest เป็นที่รู้จักในด้านความเรียบง่ายและมีประสิทธิภาพ นอกจากนี้ Random Forest เป็นเทคนิคการจำแนกข้อมูลที่ได้รับการยอมรับมากที่สุด เนื่องจากมีคุณสมบัติที่ดี เช่น การวัดความสำคัญตัวแปร ข้อผิดพลาด การวัดความคลาดเคลื่อน ความใกล้เคียง เป็นต้น [22] และ Random Forests ยังใช้ได้กับข้อมูลที่มีโครงสร้างและข้อมูลที่ไม่มีโครงสร้าง สรุปคือ Random Forests โดยทั่วไปประกอบด้วย Decision Trees หลายต้น ซึ่งเป็นการจำแนกประเภทพื้นฐานที่ประกอบขึ้นเป็น Random Forests

สมการดังนี้ [23]

$$\hat{y}_i = \sum_{k=1}^{n} f_k(x_i), \quad f_k \in F \tag{2}$$

เมื่อ $\hat{y}_i$ เป็นค่าทำนายของ i ข้อมูลที่เก็บอยู่ใน (instance) x_i F เป็นพื้นที่ของต้นไม้ถดถอย f_k สอดคล้องกับต้นไม้ และ $f_k(x_i)$ เป็นผลลัพธ์ของต้นไม้ k

4. สถิติที่ใช้ในการวัดประสิทธิภาพ

ได้แก่ ค่าความแม่นยำ (Precision) ค่าความครบถ้วน (Recall) F1-Score และค่าความถูกต้อง (Accuracy) ดังนี้

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

$$F1 = 2\,x\,\frac{\text{Precision}\ x\ \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{6}$$

เมื่อ TP (True Positive) คือ ข้อมูลที่พยากรณ์ ใช่ เทียบกับผลเฉลยคือ ใช่
TN (True Negative) คือ ข้อมูลที่พยากรณ์ ไม่ เทียบกับผลเฉลยคือ ไม่
FN (False Negative) คือ ข้อมูลที่พยากรณ์ ไม่ เทียบกับผลเฉลยคือ ใช่
FP (False Positive) คือ ข้อมูลที่พยากรณ์ ใช่ เทียบกับผลเฉลยคือ ไม่

ผลการวิจัย

1. ผลการจำแนกข้อมูลผลกระทบโควิด-19 ต่อผู้ป่วยมะเร็งตับ

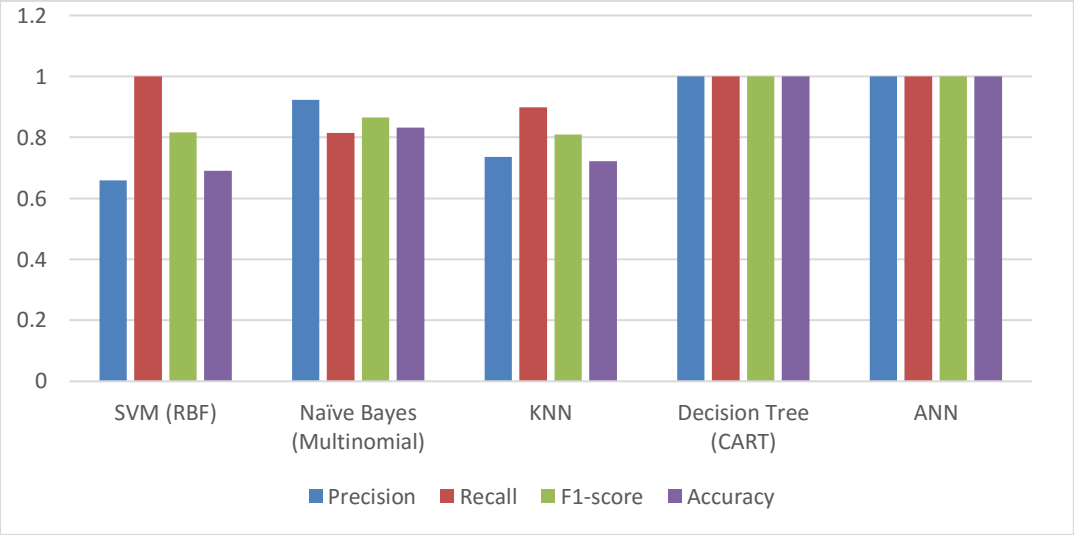
1.1 ผลการวัดประสิทธิภาพ จากการแบ่งชุดข้อมูล 80:20

ผู้วิจัยได้ดำเนินการจำแนกข้อมูลด้วยอัลกอริทึม ประกอบด้วย 1) ซัพพอร์ตเวกเตอร์แมทชีน 2) แบบนาอีฟเบย์ 3) เพื่อนบ้านใกล้ที่สุด 4) ต้นไม้ตัดสินใจ 5) โครงข่ายประสาทเทียม โดยนำข้อมูลผลกระทบโควิด-19 ต่อผู้ป่วยมะเร็งตับ [3] มาวิเคราะห์ โดยแบ่งชุดข้อมูลในการทดสอบประสิทธิภาพ 80:20 คือ แบ่งเป็นชุดข้อมูลสำหรับการเรียนรู้ 80% และชุดข้อมูลสำหรับทดสอบ 20% ผลการทดสอบประสิทธิภาพการจำแนกข้อมูล ดังตารางที่ 2 และแสดงดังภาพที่ 3

ตารางที่ 2 ผลการทดสอบประสิทธิภาพการจำแนกข้อมูล ด้วยการแบ่งชุดข้อมูล 80:20

อัลกอริทึม	Precision	Recall	F1-score	Accuracy
SVM (RBF)	0.659	1	0.792	0.656
Naïve Bayes (Multinomial)	0.923	0.814	0.865	0.833
KNN	0.736	0.898	0.809	0.722
Decision Tree (CART)	1	1	1	1
ANN	1	1	1	1

จากตารางที่ 2 การทดสอบประสิทธิภาพการจำแนกข้อมูล จากการแบ่งชุดข้อมูล 80:20 พบว่า อัลกอริทึมที่มีประสิทธิภาพมากที่สุด มีความถูกต้อง (Accuracy) เท่ากับ 100% ได้แก่ ต้นไม้ตัดสินใจ (Decision Tree) และโครงข่ายประสาทเทียม (ANN) รองลงมาได้แก่ นาอีฟเบย์ (Naïve Bayes) มีความถูกต้อง 83.3% เพื่อนบ้านใกล้ที่สุด (KNN) มีความถูกต้อง 72.2% และซัพพอร์ตเวคเตอร์แมชชีน (SVM) มีความถูกต้อง 65.6% ตามลำดับ



ภาพที่ 3 เปรียบเทียบประสิทธิภาพการจำแนกข้อมูล แบ่งชุดข้อมูล 80:20

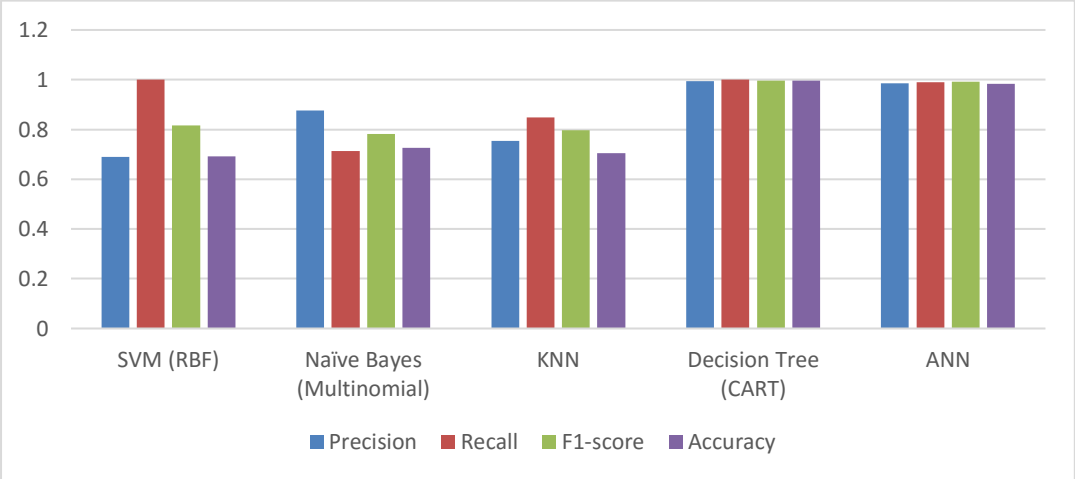
1.2 ผลการวัดประสิทธิภาพ จากการแบ่งชุดข้อมูล 5-fold cross validation

การวัดประสิทธิภาพด้วยการแบ่งข้อมูลออกเป็น 5 ชุดเท่ากัน (5-fold cross validation) เพื่อใช้ในการทดสอบความถูกต้อง ดังตารางที่ 3 และแสดงดังภาพที่ 4

ตารางที่ 3 ผลการทดสอบประสิทธิภาพการจำแนกข้อมูล ด้วยการแบ่งชุดข้อมูล 5-fold cross validation

วิธีการ	Precision	Recall	F1-score	Accuracy
SVM (RBF)	0.690	1	0.817	0.691
Naïve Bayes (Multinomial)	0.876	0.713	0.782	0.727
KNN	0.771	0.858	0.812	0.727
Decision Tree (CART)	0.994	1	0.997	0.996
ANN	0.985	0.99	0.992	0.984

จากตารางที่ 3 การทดสอบประสิทธิภาพการจำแนกข้อมูล จากการแบ่งชุดข้อมูล 5-fold cross validation พบว่า อัลกอริทึมที่มีประสิทธิภาพมากที่สุด ได้แก่ ต้นไม้ตัดสินใจ (Decision Tree) มีความถูกต้อง 99.6% รองลงมาได้แก่ โครงข่ายประสาทเทียม (ANN) มีความถูกต้อง 98.4% นาอีฟเบย์ (Naïve Bayes) และเพื่อนบ้านใกล้ที่สุด (KNN) มีความถูกต้องเท่ากัน 72.7% และซัพพอร์ตเวคเตอร์แมชชีน (SVM) มีความถูกต้อง น้อยที่สุด 69.1% ตามลำดับ



ภาพที่ 4 เปรียบเทียบประสิทธิภาพการจำแนกข้อมูล แบ่งชุดข้อมูล 5-fold cross validation

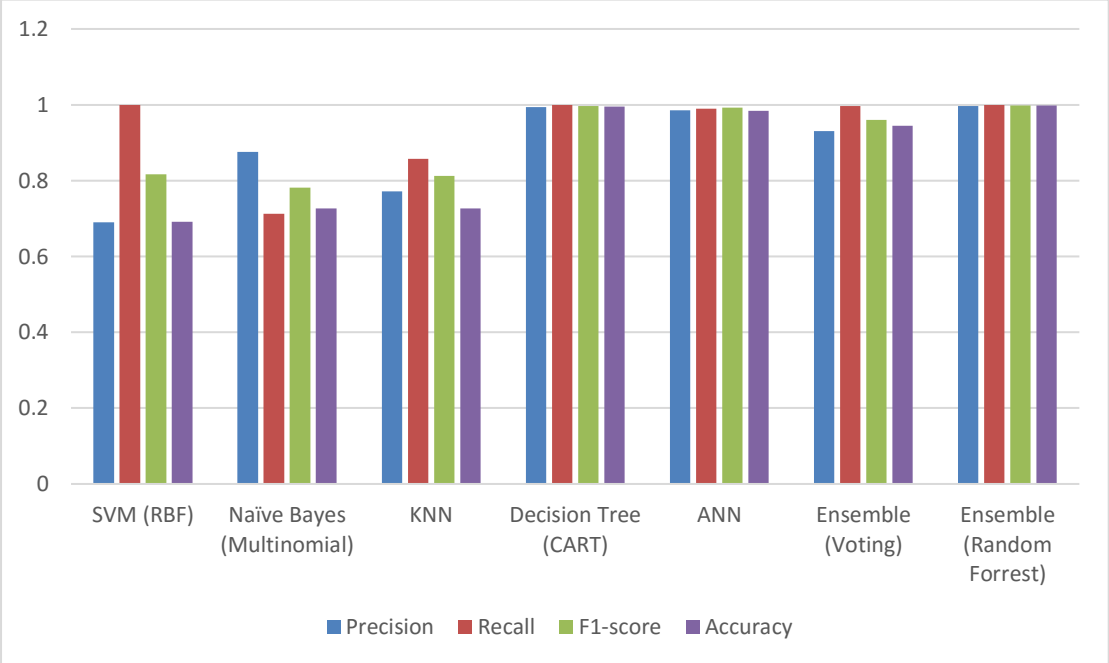
2. ผลการเพิ่มประสิทธิภาพการจำแนกข้อมูลด้วยการเรียนรู้แบบรวมกลุ่มข้อมูลผลกระทบโควิด-19 ต่อผู้ป่วยมะเร็งตับ

ผู้วิจัยได้ดำเนินการเพิ่มประสิทธิภาพจำแนกข้อมูลด้วยการสร้างโมเดลการเรียนรู้แบบรวมกลุ่ม (Ensemble) ประกอบด้วย 1) Vote Ensemble 2) Random Forest โดยนำข้อมูลผลกระทบโควิด-19 ต่อผู้ป่วยมะเร็งตับ [3] มาวิเคราะห์ โดยแบ่งชุดข้อมูลในการทดสอบประสิทธิภาพเป็น 5 ชุดเท่ากัน (5-fold cross validation) ทั้งนี้ เพื่อให้การสุ่มข้อมูลมีการกระจายตัวที่ดีที่สุด (randomization) เป็นการลด Bias จากการสุ่มข้อมูล และเพิ่มความเชื่อมั่นในการทดสอบประสิทธิภาพโมเดล ผลการทดสอบดังตารางที่ 4 และแสดงดังภาพที่ 5

ตารางที่ 4 ผลการทดสอบประสิทธิภาพการจำแนกข้อมูล

วิธีการ	Precision	Recall	F1-score	Accuracy
SVM (RBF)	0.690	1	0.817	0.691
Naïve Bayes (Multinomial)	0.876	0.713	0.782	0.727
KNN	0.771	0.858	0.812	0.727
Decision Tree (CART)	0.994	1	0.997	0.996
ANN	0.985	0.99	0.992	0.984
Ensemble (Voting)	**0.931**	**0.997**	**0.96**	**0.944**
Ensemble (Random Forest)	**0.997**	**1**	**0.998**	**0.998**

จากตารางที่ 4 การทดสอบประสิทธิภาพการจำแนกข้อมูลด้วยการเรียนรู้แบบรวมกลุ่ม จากการแบ่งชุดข้อมูล 5-fold cross validation พบว่า อัลกอริทึมที่มีประสิทธิภาพมากที่สุด ได้แก่ การสุ่มป่าไม้ (Random Forrest) มีความถูกต้อง 99.8% รองลงมาได้แก่ การจำแนกข้อมูลด้วยการโหวต (Vote Ensemble) มีความถูกต้อง 94.4% และเมื่อเปรียบเทียบกับการจำแนกข้อมูลด้วยวิธีการ SVM (RBF), Naïve Bayes (Multinomial), KNN, Decision Tree (CART), ANN พบว่า วิธีการ Ensemble (Random Forest) มีประสิทธิภาพมากกว่า



ภาพที่ 5 เปรียบเทียบการเพิ่มประสิทธิภาพการจำแนกข้อมูลด้วยการเรียนรู้แบบรวมกลุ่ม

อภิปรายผลการวิจัย

1. ผลการจำแนกข้อมูลผลกระทบโควิด-19 ต่อผู้ป่วยมะเร็งตับ

1.1 ผลการวัดประสิทธิภาพ จากการแบ่งชุดข้อมูล 80:20 พบว่า อัลกอริทึมที่มีประสิทธิภาพมากที่สุด ได้แก่ ต้นไม้ตัดสินใจ (Decision Tree) และโครงข่ายประสาทเทียม (ANN) มีความถูกต้อง 100% รองลงมาได้แก่ นาอีฟเบย์ (Naïve Bayes) มีความถูกต้อง 83.3% เพื่อนบ้านใกล้ที่สุด (KNN) มีความถูกต้อง 72.2% และซัพพอร์ตเวคเตอร์แมชชีน (SVM) มีความถูกต้อง 65.6% ทั้งนี้เนื่องจาก อัลกอริทึมต้นไม้ตัดสินใจ ด้วยวิธีการ CART ได้มีการจัดการเตรียมข้อมูลเป็นตัวเลข และมีลักษณะข้อมูลแบบไบนารีเป็นส่วนมาก จึงส่งผลให้มีประสิทธิภาพมาก สอดคล้องกับงานวิจัยของ Patel และ Prajapati [10] ได้ศึกษาการจำแนกข้อมูลด้วยต้นไม้ตัดสินใจ ชุดข้อมูลการทำลองเป็นชุดข้อมูลรถยนต์ ทำการเปรียบเทียบอัลกอริทึม ID3, C4.5 CART พบว่า อัลกอริทึม CART มีประสิทธิภาพมากที่สุด มีความถูกต้อง 97.11% ส่วนอัลกอริทึมโครงข่ายประสาทเทียม มีผลการวัดประสิทธิภาพเท่ากัน ทั้งนี้เนื่องจากผู้วิจัยได้มีการกำหนดการเรียนรู้แบบหลายชั้น จึงส่งผลให้มีประสิทธิภาพเท่ากับต้นไม้ตัดสินใจ แต่ทั้งนี้การทดสอบเป็นแบบแบ่งชุดข้อมูล 80:20 อาจจะมีการสุ่มข้อมูลทดสอบง่าย จึงส่งผลให้มีประสิทธิภาพมากตามไปด้วย

1.2 ผลการวัดประสิทธิภาพการจำแนกข้อมูล จากการแบ่งชุดข้อมูล 5-fold cross validation พบว่า อัลกอริทึมที่มีประสิทธิภาพมากที่สุด ได้แก่ ต้นไม้ตัดสินใจ (Decision Tree) มีความถูกต้อง 99.6% รองลงมาได้แก่ โครงข่ายประสาทเทียม (ANN) มีความถูกต้อง 98.4% นาอีฟเบย์ (Naïve Bayes) และเพื่อนบ้านใกล้ที่สุด (KNN) มีความถูกต้องเท่ากัน 72.7% และซัพพอร์ตเวคเตอร์แมชชีน (SVM) มีความถูกต้อง น้อยที่สุด 69.1% ทั้งนี้เนื่องจากอัลกอริทึมต้นไม้ตัดสินใจ เลือกวิธีการแบบ CART ได้มีการจัดการเตรียมข้อมูลเป็นตัวเลข และมีลักษณะข้อมูลแบบไบนารีเป็นส่วนมาก จึงส่งผลให้มีประสิทธิภาพมาก สอดคล้องกับงานวิจัยของ Patel และ Prajapati [10] ได้ศึกษาการจำแนกข้อมูลด้วยต้นไม้ตัดสินใจ ชุดข้อมูลการทำลองเป็นชุดข้อมูลรถยนต์ พบว่า อัลกอริทึม CART มีประสิทธิภาพมากที่สุด มีความถูกต้อง 97.11% เมื่อเทียบกับวิธีการอื่น

จากการประเมินสิทธิภาพด้วยวิธีการแบ่งชุดข้อมูลเป็น 2 ส่วน (Split Test) คือ 80:20 และวิธีการแบบแบ่งชุดข้อมูล 5-fold cross validation พบว่า การแบ่งชุดข้อมูล 2 ส่วนมีประสิทธิภาพมากกว่า เนื่องจากการแบ่งชุดข้อมูลแบบ Split Test เป็นการสุ่มข้อมูลกระจายไปในชุดข้อมูลการเรียนรู้ (Train) จำนวนมากกว่าการแบ่งชุดข้อมูล 5-fold cross validation แต่ก็อาจมี Bias เนื่องจากมีการสุ่มข้อมูลครั้งเดียว ทำให้ต้องมีการทดสอบประสิทธิภาพด้วยวิธีการแบบ k-fold cross validation เพื่อให้การสุ่มข้อมูลมีการกระจายตัวที่ดีที่สุด (randomization) เป็นการลด Bias จากการสุ่มข้อมูล และเพิ่มความเชื่อมั่นในการทดสอบประสิทธิภาพโมเดล ซึ่งผลการการทดสอบประสิทธิภาพมีการแบ่งชุดข้อมูล 5-fold cross validation คือแบ่งชุดข้อมูลเท่ากันจำนวน 5 ชุด ต้องทำการเทรนโมเดลทั้งหมด 5 รอบ และทดสอบโมเดล 5 รอบ ซึ่งการเทรนโมเดลการเรียนรู้ของเครื่องจะมีข้อมูลน้อยกว่าการแบ่งชุดข้อมูลแบบ Split Test ส่งผลให้ประสิทธิภาพลดลงตามไปด้วย

2. ผลการเพิ่มประสิทธิภาพการจำแนกข้อมูลด้วยการเรียนรู้แบบรวมกลุ่มข้อมูลผลกระทบโควิด-19 ต่อผู้ป่วยมะเร็งตับ

ผลการเพิ่มประสิทธิภาพการจำแนกข้อมูลด้วยการเรียนรู้แบบรวมกลุ่ม (Ensemble) จากการแบ่งชุดข้อมูล 5-fold cross validation พบว่า อัลกอริทึมที่มีประสิทธิภาพมากที่สุด ได้แก่ การสุ่มป่าไม้ (Random Forrest) มีความถูกต้อง 99.8% รองลงมาได้แก่ การจำแนกข้อมูลด้วยการโหวต (Vote Ensemble) มีความถูกต้อง 94.4% ทั้งนี้ เนื่องจาก Random Forrest เป็นการจำแนกข้อมูลประเภทจากโครงสร้างต้นไม้ เป็น Decision Trees หลายต้น ประกอบด้วยอัลกอริทึมเฉลี่ยสองอัลกอรีทึมของต้นไม้ตัดสินใจแบบสุ่ม คือ อัลกอริทึม Random Forest และวิธีการ Extra-Trees ซึ่งทั้งสองวิธีเป็นเทคนิคการผสมผสาน นอกจากนี้ Random Forrest มีคุณสมบัติที่ดี เช่น การวัดความสำคัญตัวแปร ข้อผิดพลาด การวัดความคลาดเคลื่อน ความใกล้เคียง ส่งผลให้สามารถเพิ่มประสิทธิภาพการจำแนกข้อมูล เมื่อเปรียบเทียบจากอัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) มีความถูกต้อง 99.6% ส่วน Random Forrest มีความถูกต้อง 99.8% ดังนั้น จะเห็นได้ว่าการจำแนกข้อมูลมีประสิทธิภาพมากขึ้น โดยมีความถูกต้องเพิ่มขึ้น 0.2% สอดคล้องกับงานวิจัยของ Mohana และคนอื่น [12] ได้ศึกษาอัลกอริทึมการสุ่มป่าไม้ (Random Forest) สำหรับการจำแนกตามกลุ่มต้นไม้ (tree based ensemble) ด้วยวิธีการ Classification by ensembles from Random partitions (CERP) ข้อมูลเป็นชุดข้อมูลมะเร็งต่อมลูกหมาก ผ่านกระบวนการเรียนรู้ของเครื่อง ด้วยกระบวนการตรวจสอบข้อมูลและทำความสะอาดข้อมูล ผลการศึกษาพบว่าการจำแนกข้อมูลด้วยวิธีการ CERP มีความถูกต้อง 0.8333

ข้อเสนอแนะ

การเพิ่มประสิทธิภาพจำแนกข้อมูลผลกระทบโควิด-19 ต่อผู้ป่วยมะเร็งตับ สามารถนำไปใช้กับผู้ป่วยมะเร็งตับที่ได้รับผลกระทบต่อการติดเชื้อโควิด-19 เพื่อให้โรงพยาบาลหรือสาธารณสุข สามารถนำโมเดลการจำแนกข้อมูลผู้ป่วยมะเร็งตับได้อย่างรวดเร็ว มีความถูกต้องมากยิ่งขึ้น รวมทั้งพัฒนาต่อโดยนำไปใช้เป็นข้อมูลสำหรับการพยากรณ์ผู้มีความเสี่ยงเพิ่มเติมในอนาคต และการวิจัยครั้งต่อไปควรหาวิธีการให้ค่าความแม่นยำสูงมากกว่านี้ เช่นวิธีการ Deep Learning วิธีการจำแนกข้อมูลการเรียนรู้แบบรวมกลุ่มด้วยเทคนิคการผสมผสาน เป็นต้น

เอกสารอ้างอิง

[1] Chan, S. L., & Kudo, M. (2020). Impacts of COVID-19 on Liver Cancers: During and after the Pandemic. *Liver Cancer*, 9(5), 491-502. doi : https://doi.org/10.1159/000510765

[2] Geh, D., Watson, R., Sen, G., French, J. J., Hammond, J., Turner, P., ... & Reeves, H. L. (2022). COVID-19 and liver cancer: lost patients and larger tumours. *BMJ open gastroenterology*, 9(1), doi: 10.1136/bmjgast-2021-000794

[3] fedesoriano. (September 2022). COVID-19 effect on Liver Cancer Prediction Dataset. Retrieved [30 September 2022] from https://www.kaggle.com/datasets/fedesoriano/covid19-effect-on-liver-cancer-prediction-dataset.

[4] Ghosh, S., Dasgupta, A., & Swetapadma, A. (2019, February). A study on support vector machine based linear and non-linear pattern classification. In *2019 International Conference on Intelligent Sustainable Systems (ICISS),* 24-28, IEEE.

[5] Mohan, L., Pant, J., Suyal, P., & Kumar, A. (2020, September). Support vector machine accuracy improvement with classification. In *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN),* 477-481, IEEE.

[6] Harahap, F., Harahap, A. Y. N., Ekadiansyah, E., Sari, R. N., Adawiyah, R., & Harahap, C. B. (2018, August). Implementation of Naïve Bayes classification method for predicting purchase. In *2018 6th International Conference on Cyber and IT Service Management (CITSM),* 1-5, IEEE.

[7] Singh, G., Kumar, B., Gaur, L., & Tyagi, A. (2019, April). Comparison between multinomial and Bernoulli naïve Bayes for text classification. In *2019 International Conference on Automation, Computational and Technology Management (ICACTM)*, 593-596, IEEE.

[8] Abbas, M., Memon, K. A., Jamali, A. A., Memon, S., & Ahmed, A. (2019). Multinomial Naive Bayes classification model for sentiment analysis. *IJCSNS International Journal of Computer Science and Network Security*, 19(3), 62-67, doi: 10.13140/RG.2.2.30021.40169

[9] Xing, W., & Bei, Y. (2019r). Medical health big data classification based on KNN classification algorithm. *IEEE Access*, 8, 28808-28819. doi: 0.1109/ACCESS.2019.2955754

[10] Patel, H. H., & Prajapati, P. (2018). Study and analysis of decision tree based classification algorithms. *International Journal of Computer Sciences and Engineering*, 6(10), 74-78. doi: 10.26438/ijcse/v6i10.7478

[11] Al-Massri, R., Al-Astel, Y., Ziadia, H., Mousa, D. K., & Abu-Naser, S. S. (2018). Classification Prediction of SBRCTs Cancers Using Artificial Neural Network. *International journal of Academic Engineering Research (IJAER),* 2(11), 1-7.

[12] Mohana, R. M., Reddy, C. K. K., Anisha, P. R., & Murthy, B. R. (2021). Random forest algorithms for the classification of tree-based ensemble. *Materials Today: Proceedings*, 1-6. https://doi.org/10.1016/j.matpr.2021.01.788

[13] Zhang, S. (2020). Cost-sensitive KNN classification. *Neurocomputing*, 391, 234-242. doi : https://doi.org/10.1016/j.neucom.2018.11.101

[14] Abdulkareem, N. M., & Abdulazeez, A. M. (2021). Machine learning classification based on Radom Forest Algorithm: A review. *International Journal of Science and Business*, 5(2), 128-142. doi: https://doi.org/10.5281/zenodo.4471118

[15] Li, Y., Jiang, Z. L., Yao, L., Wang, X., Yiu, S. M., & Huang, Z. (2019). Outsourced privacy-preserving C4.5 decision tree algorithm over horizontally and vertically partitioned dataset among multiple parties. *Cluster Computing,* 22(S1), 1581–1593. doi: https://doi.org/10.1007/s10586-017-1019-9

[16] Singh, S., & Gupta, P. (2014). Comparative study ID3, cart and C4. 5 decision tree algorithm: a survey. *International Journal of Advanced Information Science and Technology (IJAIST),* 27(27), 97-103. doi:10.15693/ijaist/2014.v3i7.47-52

[17] Band, S. S., Janizadeh, S., Saha, S., Mukherjee, K., Bozchaloei, S. K., Cerdà, A., Shokri, M., & Mosavi, A. (2020). Evaluating the Efficiency of Different Regression, Decision Tree, and Bayesian Machine Learning Algorithms in Spatial Piping Erosion Susceptibility Using ALOS/PALSAR Data. *Land*, 9(10), 346. doi: https://doi.org/10.3390/land9100346

[18] Sarker, I. H., Colman, A., Han, J., Khan, A. I., Abushark, Y. B., & Salah, K. (2020). BehavDT: A Behavioral Decision Tree Learning to Build User-Centric Context-Aware Predictive Model. *Mobile Networks and Applications*, 25(3), 1151–1161. doi: https://doi.org/10.1007/s11036-019-01443-z

[19] Alghoul, A., Al Ajrami, S., Al Jarousha, G., Harb, G., & Abu-Naser, S. S. (2018). Email classification using artificial neural network. *International journal of Academic Engineering Research (IJAER)*, 2(11), 8-14.

[20] ABRO, A. A. (2021). Vote-based: Ensemble approach. Sakarya University *Journal of Science*, 25(3), 858-866. doi: https://doi.org/10.16984/saufenbilder.901960

[21] Reis, I., Baron, D., & Shahaf, S. (2018). Probabilistic Random Forest: A Machine Learning Algorithm for Noisy Data Sets. *The Astronomical Journal*, 157(1), 1-12. doi: https://doi.org/10.3847/1538-3881/aaf101

[22] Speiser, J.L., Miller, M.E., Tooze, J., Edward I., (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems With Applications*, 134, 93-101. doi: https://doi.org/10.1016/j.eswa.2019.05.028

[23] Wang, Y., Pan, Z., Zheng, J., Qian, L., & Li, M. (2019). A hybrid ensemble method for pulsar candidate classification. *Astrophysics and Space Science,* 364(8), 1-13. doi: 10.1007/s10509-019-3602-4