บทที่ 3 วิธีดำเนินการวิจัย

3.1 เครื่องมือดำเนินการวิจัย

1. Python
2. Visual studio code
3. Google Colab
4. Chemdoodle
5. Microsoft excel
6. Microsoft word

3.2 วิธีการดำเนินงาน

1) การเก็บรวบรวมข้อมูล

รวบรวมงานวิจัยที่ใช้สร้างแบบจำลองเพื่อทำนายความเป็นพิษทางผิวหนังใช้ข้อมูลจากงานวิจัยของ Srisongkram โดยมีสารเคมีทั้งหมด 469 ตัวและมีค่า Potential of Half maximal inhibitory concentration (pIC50) ของสารเคมีแต่ละตัวที่ถูกทดสอบกับเซลล์เคราติโนไซต์อมตะของมนุษย์ (HaCaT) และใช้ข้อมูลจากงานวิจัยของ Han et al. (2021) เพื่อทดสอบความถูกต้องของแบบจำลองโดยใช้ข้อมูลในส่วนที่เป็นการทดสอบการระคายเคืองผิวหนังในสัตว์ทดลอง (In vivo) เป็นตัวทดสอบ

2) การปรับแต่งข้อมูล

เนื่องจากข้อมูลในงานวิจัยของ Srisongkram ส่วนของค่า negative log cell viability of Half maximal inhibitory concentration (pIC50) ของสารเคมีแต่ละตัวไม่สามารถนำมาใช้เพื่อสร้างแบบจำลองที่ต้องการจึงมีการแปลงค่าส่วนนี้เป็นข้อมูลรูปแบบ IC50 และสุดท้ายจะถูกแปลงเป็นข้อมูลเชิงคุณภาพโดยหากค่า IC50 สูงกว่า 10 ไมโครโมลาร์จะถูกนับเป็นสารเคมีที่ไม่ก่อให้เกิดการระคายเคืองผิวหนัง หากนอกเหนือจากนั้นจะเป็นสารเคมีที่ก่อให้เกิดการระคายเคืองผิวหนัง และข้อมูลจากงานวิจัยของ Han et al. (2021) ต้องแปลงโครงสร้างของสารเคมีแต่ละตัวเป็น SMILEs และมีการตัดข้อมูลที่เป็นสารผสม (Mixture) ข้อมูลที่เป็นสารเคมีอนินทรีย์ (Inorganic) และข้อมูลซ้ำออกจากฐานข้อมูลด้วย

3) การคำนวณคุณสมบัติทางเคมีกายภาพ

มีการคำนวณคุณสมบัติทางเคมีกายภาพคือ LogP มวลโมเลกุล Hydrogen bond donor (HBD) กับ Hydrogen bond acceptor (HBA) เพื่อวิเคราะห์ว่าสามารถใช้คุณสมบัติทางเคมีกายภาพในทำนายความเป็นพิษทางผิวหนังของสารเคมีได้หรือไม่ ด้วยการใช้อัลกอลิทึมที่เขียนด้วยภาษา Python

4) การคำนวณลายพิมพ์ระดับโมเลกุล (Molecular Fingerprints)

Canonical Simplified molecular-input line-entry system (SMILEs) จะถูกนำมาใช้เป็นข้อมูลโครงสร้างของสารเคมีก่อนเนื่องจากสามารถอธิบายโครงสร้างโมเลกุลให้อยู่ในรูปแบบข้อมูลและรูปแบบข้อมูลชุดนี้จะไม่ซ้ำกันในแต่ละสารเคมีด้วยการแปลงมาจาก SMILEs จากนั้นคำนวณลายพิมพ์ระดับโมเลกุล (Molecular Fingerprints) ทั้งในรูปแบบ Pubchem และ Substructure ด้วยข้อมูล Canonical SMILEs ของสารเคมีโดยใช้อัลกอลิทึมที่เขียนด้วยภาษา Python นำไปพัฒนาแบบจำลองเพื่อใช้ในการทำนายความเป็นพิษทางผิวหนังด้วยการใช้การเรียนรู้ของเครื่อง

5) สร้างกราฟการวิเคราะห์เชิงองค์ประกอบหลัก (Principal component analysis)

การวิเคราะห์เชิงองค์ประกอบหลัก (PCA) คือวิธีการลดมิติของข้อมูลลงเพื่อให้สามารถสำรวจวิเคราะห์ข้อมูลได้ง่ายยิ่งขึ้น โดยจะใช้ลายพิมพ์ระดับโมเลกุล (Molecular Fingerprints) ทั้งในรูปแบบ Pubchem และ Substructure เพื่อสร้างกราฟการวิเคราะห์เชิงองค์ประกอบหลักด้วยการใช้อัลกอลิทึมที่เขียนด้วยภาษา Python ซึ่งบ่งชี้ความสามารถในการแยกสารเคมีที่เป็นพิษทางผิวหนังออกจากสารเคมีที่ไม่เป็นพิษทางผิวหนังด้วย PC1 กับ PC2

6) สร้างและออกแบบแบบจำลอง

การสร้างแบบจำลองด้วยการใช้เทคนิคการหาความสัมพันธ์เชิงโครงสร้างและการออกฤทธิ์ของสารเคมี (Quantitative Structure Activity Relationship; QSAR) ร่วมกับการเทคนิค แรนดอมฟอเรส (Random Forest) ซึ่งเป็นอัลกอริทึมการเรียนรู้เองของคอมพิวเตอร์ (machine learning) จะเริ่มจากการสืบค้นชุดข้อมูลที่เกี่ยวข้องแล้วนำข้อมูลที่ได้มานั้นตัดค่าที่ไม่มีความจำเป็นออก หลังจากนั้นทำการแบ่งชุดข้อมูลออกเป็น 2 ชุด คือข้อมูลชุดเรียนรู้ (training set) มีทั้งสิ้น 328 ตัว และข้อมูลชุดทดสอบ (Test set) มีทั้งสิ้น 141 ตัว โดยการสร้างแบบจำลองเพื่อประเมินแบบจัดหมวดหมู่ (Classification-based model) จะถูกใช้ในโครงการนี้

7) ประเมินและตรวจทานแบบจำลอง

การทดสอบภาวะสารูปสนิทดี (Goodness-of-fit test) จะถูกใช้ในการประเมินแบบจำลองการศึกษาความสัมพันธ์ระหว่างโครงสร้างและความเป็นพิษทางผิวหนังของสารด้วยการเรียนรู้ของเครื่อง (OECD, 2014) โดยใช้ Confusion matrix ในประเมินด้านความไว (Sensitivity) ความจำเพาะ (Specificity) ความแม่นยำ (Accuracy) และความเที่ยง (Precision) และมีตัวแปร 4 ตัวที่เกี่ยวข้องกับการคำนวณ (Bank & Schmehl, 1989; Wang et al., 2010) ได้แก่

1) True positive (TP) คือ สิ่งที่ทำนายตรงกับสิ่งที่เกิดขึ้นจริง กรณีทำนายได้ผลบวก สิ่งที่เกิดขึ้นให้ผลบวก
2) True negative (TN) คือ สิ่งที่ทำนายตรงกับสิ่งที่เกิดขึ้นจริง กรณีทำนายได้ผลลบ สิ่งที่เกิดขึ้นให้ผลลบ
3) False positive (FP) คือ สิ่งที่ทำนายไม่ตรงกับสิ่งที่เกิดขึ้นจริง กรณีทำนายได้ผลบวก สิ่งที่เกิดขึ้นให้ผลลบ
4) False negative (FN) คือ สิ่งที่ทำนายไม่ตรงกับสิ่งที่เกิดขึ้นจริง กรณีทำนายได้ผลลบ สิ่งที่เกิดขึ้นให้ผลบวก

Confusion matrix	สิ่งที่เกิดขึ้นให้ผลบวก	สิ่งที่เกิดขึ้นให้ผลลบ
ทำนายได้ผลบวก	True positive	False positive
ทำนายได้ผลลบ	False negative	True negative

ตารางที่ 2 ตารางแสดง Confusion matrix

ความไว (sensitivity) คือ สัดส่วนผลบวกที่เป็นจริง ใช้แยกผลลบที่ไม่เป็นจริงออกเนื่องจากยิ่งความไวมากเท่าใด โอกาสได้ผลลบที่ไม่เป็นจริงยิ่งน้อยลงเท่านั้น คำนวณได้จาก Sensitivity = TP/(TP + FN)

ความจำเพาะ (specificity) คือ สัดส่วนผลลบที่เป็นจริง ใช้แยกผลบวกที่ไม่เป็นจริงออก ยิ่งความจำเพาะสูงเท่าใด ยิ่งสามารถแยกผลบวกที่ไม่เป็นจริงได้มากเท่านั้น คำนวณได้จาก Specificity = TN/(TN + FP)

ความถูกต้อง (Accuracy) คือ ความสามารถที่บ่งบอกว่าการทดสอบมีผลใกล้เคียงกับค่าจริง คำนวณได้จาก Accuracy = (TP + TN)/(TP+ TN + FP + FN)

ความแม่นยำ (Precision) คือ ความสามารถในการจำแนกผลบวกแท้จากผลบวกทั้งหมด คำนวณได้จาก Precision = TP/(TP + FP)

8) Receiver Operating Characteristics (ROC) Curve

Receiver Operating Characteristics (ROC) Curve ของแบบจำลองแรนดอมฟอเรส (Random Forest) ที่สร้างขึ้นจากลายพิมพ์ระดับโมเลกุลทั้งสองรูปแบบคือลายพิมพ์ระดับโมเลกุลแบบ Pubchem และแบบ Substructure ด้วยภาษา Python บ่งชี้ถึงประสิทธิภาพของความสามารถในการทำนายของแบบจำลอง

9) ตรวจสอบประสิทธิภาพของแบบจำลองใน Applicability Domain

ใช้อัลกอลิทึมที่เขียนด้วยภาษา python ในการสร้างแผนภูมิที่แสดงความสัมพันธ์ระหว่าง Matrix value กับ K value

10) ตรวจสอบลายพิมพ์ระดับโมเลกุลที่สำคัญต่อการทำนายพิษทางผิวหนังของแบบจำลอง

3.3 สถานที่ทำวิจัย

- ห้องปฏิบัติการวิจัยคณะเภสัชศาสตร์ มหาวิทยาลัยขอนแก่น

3.4 แผนการดำเนินการเกี่ยวกับกิจกรรมและระยะเวลาในการทำวิจัย (ระยะเวลา 3-6 เดือน)

แผนการดำเนินงาน	พ.ศ. 2566					พ.ศ. 2567	
	ส.ค.	ก.ย.	ต.ค.	พ.ย.	ธ.ค.	ม.ค.	ก.พ.
1.จัดทำเค้าโครง							
2.รวบรวมข้อมูลสำหรับสร้างแบบจำลอง							
3.สร้างและออกแบบแบบจำลอง							
4.ประเมินและตรวจทานแบบจำลอง							
5.วิเคราะห์ข้อมูลด้วยแบบจำลองและประเมินประสิทธิภาพของแบบจำลอง							
6.สรุปผลการศึกษาในการพัฒนาแบบจำลองการเรียนรู้ของเครื่อง จัดทำรูปเล่มและเสนอโครงงานวิจัย							

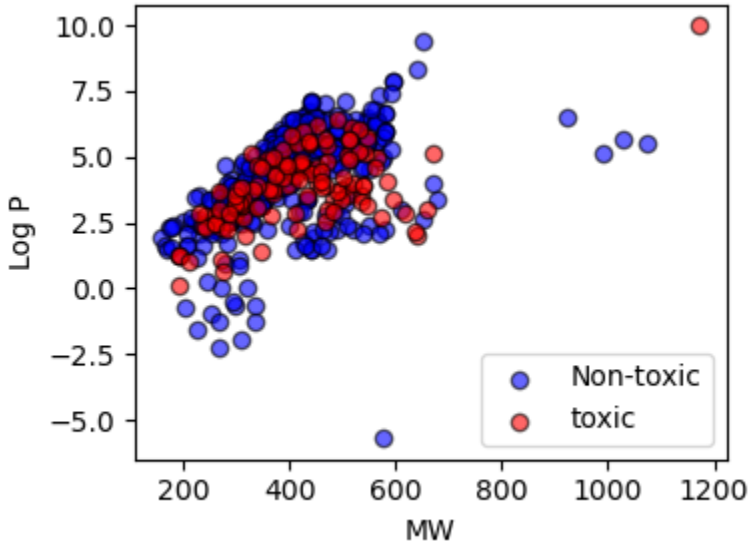
3.5 งบประมาณ

งบประมาณจัดโครงการ จากการสนับสนุนจากคณะเภสัชศาสตร์ เป็นเงินทั้งสิ้น 2,000 บาท โดยมีรายละเอียดดังนี้

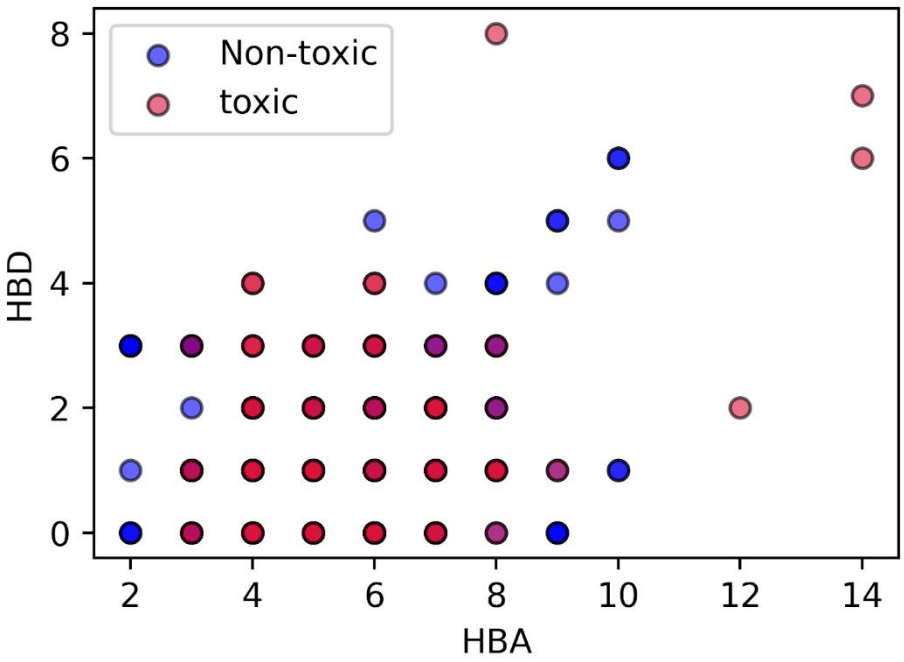
- ค่า Chemdoodle 1,500 บาท
- ค่าทำ poster และรูปเล่ม 500 บาท

บทที่ 4 ผลการทดลอง

การทดลองได้เน้นการวิเคราะห์การกระจายของสารเคมีโดยใช้คุณลักษณะทางเคมีกายภาพและลายพิมพ์ระดับโมเลกุลต่าง ๆ รูปที่ 1 แสดงการกระจายของสารเคมีจำนวน 469 ตัวด้วยการใช้ลักษณะทางเคมีกายภาพคือ LogP กับมวลโมเลกุลโดยใช้อัลกอลิทึมที่สร้างด้วยภาษา Python มีค่าเฉลี่ยมวลโมเลกุลอยู่ที่ 400.68±127.49 และมี LogP เฉลี่ยอยู่ที่ 4.0±1.79 และรูปที่ 2 แสดงการกระจายของสารเคมีจำนวน 469 ตัว ด้วยการใช้ลักษณะทางเคมีกายภาพคือ Hydrogen bond donor (HBD) กับ Hydrogen bond acceptor (HBA) โดยใช้อัลกอลิทึมที่สร้างด้วยภาษา Python พบว่ามีค่าเฉลี่ย HBD เท่ากับ 1.29±1.26 และค่าเฉลี่ย HBA เท่ากับ 5.3±2.1 ซึ่งแทนจุดข้อมูลเป็นวงกลมสีน้ำเงินสำหรับสารเคมีที่ไม่เป็นพิษและวงกลมสีแดงสำหรับสารเคมีที่เป็นพิษ จากผลลัพธ์ที่ได้บ่งชี้ว่าไม่สามารถแยกสารเคมีที่เป็นพิษต่อผิวหนังด้วยคุณสมบัติทางเคมีกายภาพ เมื่อดูจากรูปจะพบว่าสารเคมีที่เป็นพิษและไม่เป็นพิษมีคุณสมบัติทางเคมีกายภาพคล้ายคลึงกัน



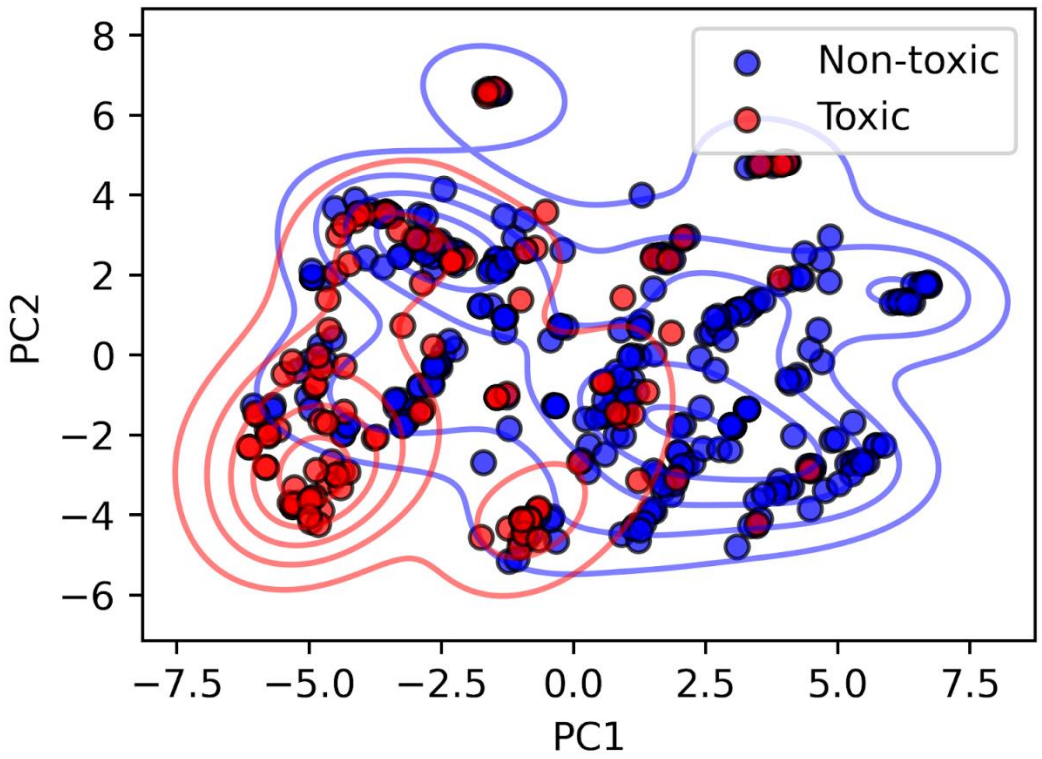
รูปที่ 1 แสดงการกระจายของโมเลกุลจำนวน 469 ตัวด้วยการใช้ลักษณะทางเคมีกายภาพคือ logP กับมวลโมเลกุล ซึ่งแทนจุดข้อมูลเป็นวงกลมสีน้ำเงินสำหรับโมเลกุลที่ไม่เป็นพิษและวงกลมสีแดงสำหรับโมเลกุลที่เป็นพิษ



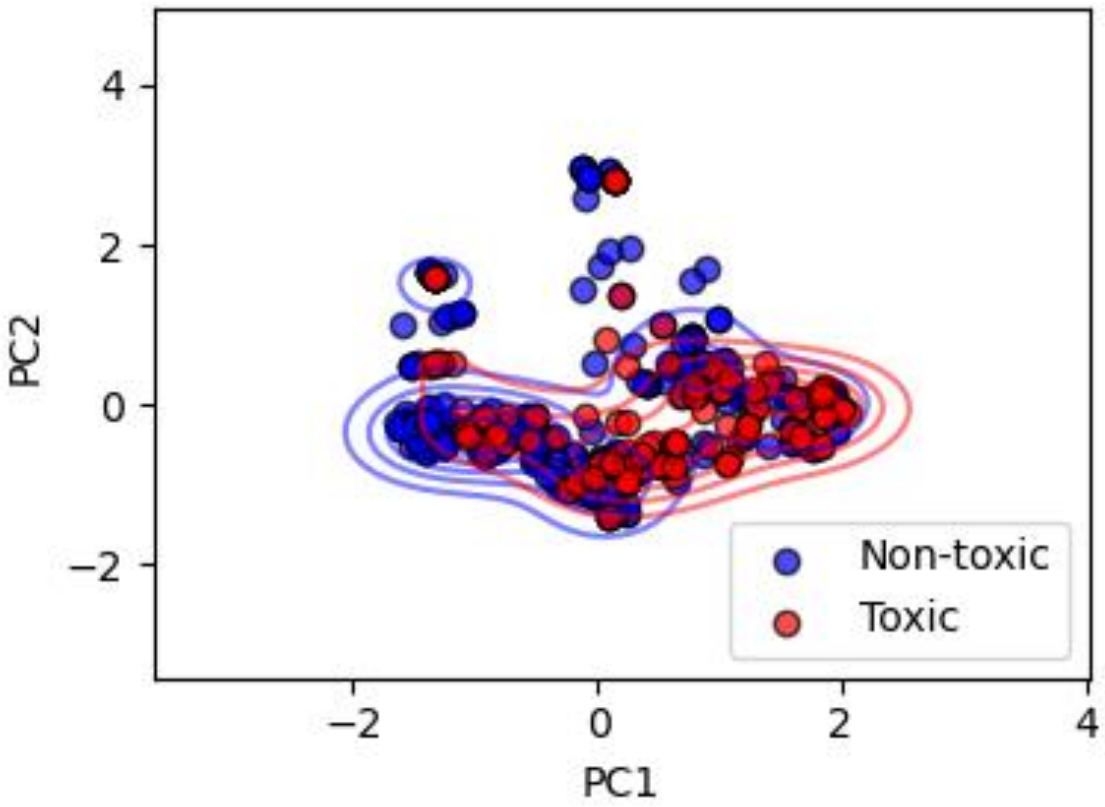
รูปที่ 2 แสดงการกระจายของสารเคมีจำนวน 469 โมเลกุลด้วยการใช้ลักษณะทางเคมีกายภาพคือ Hydrogen bond donor กับ Hydrogen bond acceptor ซึ่งแทนจุดข้อมูลเป็นวงกลมสีน้ำเงินสำหรับสารเคมีที่ไม่เป็นพิษและวงกลมสีแดงสำหรับสารเคมีที่เป็นพิษ

Principal component analysis (PCA)

ใช้อัลกอลิทึม PCA ด้วยภาษา Python เพื่อแปลงลายพิมพ์ระดับโมเลกุลของ PubChem และ Substructure ของสารเคมี 469 โมเลกุลเป็น Principal component 1 กับ 2 ซึ่งแทนจุดข้อมูลเป็นวงกลมสีน้ำเงินสำหรับสารเคมีที่ไม่เป็นพิษและวงกลมสีแดงสำหรับสารเคมีที่เป็นพิษ ดังรูปที่ 3 และ 4 ตามลำดับ จากผลลัพธ์บ่งชี้ว่าไม่สามารถแยกสารเคมีที่เป็นพิษต่อผิวหนังด้วย Principal component ได้ เมื่อดูจากรูปจะพบว่าสารเคมีที่เป็นพิษและไม่เป็นพิษมีลักษณะทาง Principal component คล้ายคลึงและซ้อนทับกันกัน



รูปที่ 3 การกระจายของสารเคมี 469 โมเลกุลด้วยการใช้ Principal component 1 กับ 2 โดยการแปลงจากลายพิมพ์ระดับโมเลกุลของ PubChem ซึ่งแทนจุดข้อมูลเป็นวงกลมสีน้ำเงินสำหรับสารเคมีที่ไม่เป็นพิษและวงกลมสีแดงสำหรับสารเคมีที่เป็นพิษ



รูปที่ 4 การกระจายของสารเคมี 469 โมเลกุลด้วยการใช้ Principal component 1 กับ 2 โดยการแปลงจากลายพิมพ์ระดับโมเลกุลของ Substructure ซึ่งแทนจุดข้อมูลเป็นวงกลมสีน้ำเงินสำหรับสารเคมีที่ไม่เป็นพิษและวงกลมสีแดงสำหรับสารเคมีที่เป็นพิษ

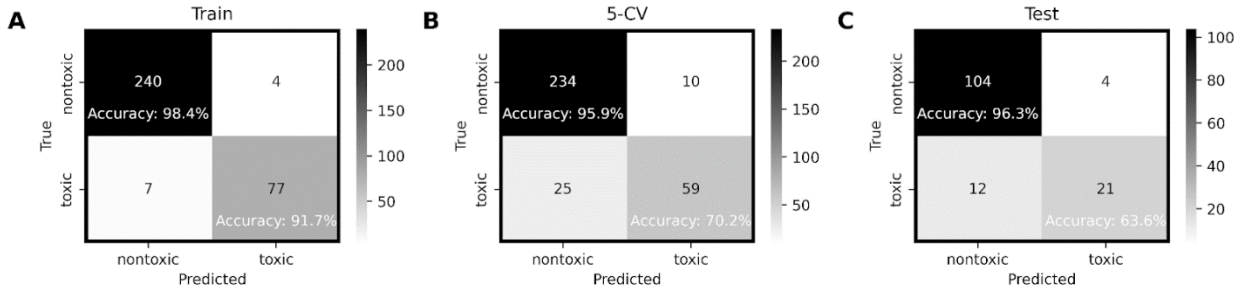
Model construction

ใช้อัลกอลิทึมสร้างแบบจำลองแรนดอมฟอเรส (Random Forest) ด้วยภาษา Python โดยให้ตัวแปรต้นเป็นลายพิมพ์ระดับโมเลกุลทั้งลายพิมพ์ระดับโมเลกุลแบบ Pubchem และแบบ Substructure ในขณะที่ตัวแปรตามเป็นความเป็นพิษทางผิวหนังเพื่อใช้ทำนายความเป็นพิษทางผิวหนังจากลายพิมพ์ระดับโมเลกุลของสารเคมีที่ต้องการ และหาประสิทธิภาพของแบบจำลองโดยดูจากค่าที่แสดงประสิทธิภาพของแบบจำลอง ได้แก่ ความถูกต้อง แม่นยำ ความไว และความจำเพาะดังตารางที่ 1 ซึ่งบ่งชี้ว่าแบบจำลองที่สร้างด้วยลายพิมพ์ระดับโมเลกุลทั้งสองมีประสิทธิภาพใกล้เคียงกันและสามารถใช้ในการทำนายความเป็นพิษทางผิวหนังของสารเคมีโดยใช้ลายพิมพ์ระดับโมเลกุลของสารเคมีนั้นได้

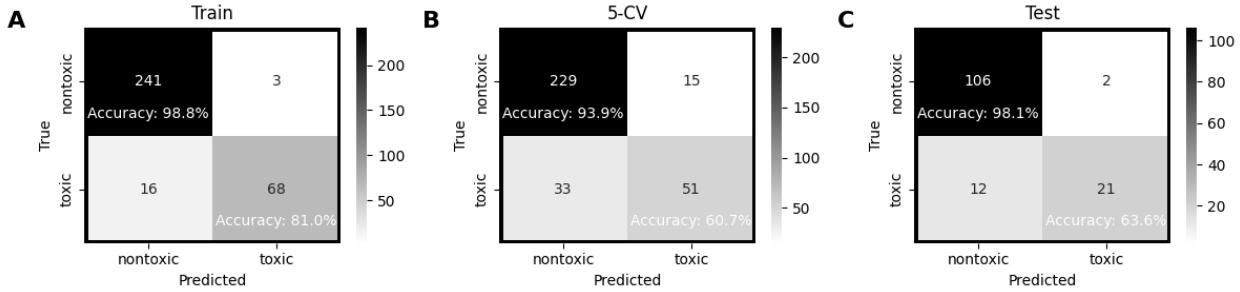
model	accuracy			precision			sensitivity			specificity		
	train	cv	test	train	cv	test	train	cv	test	train	cv	test
Pubchem	0.97	0.89	0.89	0.95	0.85	0.84	0.92	0.70	0.64	0.98	0.96	0.96
Substructure	0.94	0.85	0.90	0.96	0.77	0.91	0.81	0.61	0.64	0.99	0.94	0.98

ตารางที่ 1 แสดงประสิทธิภาพของแบบจำลองแรนดอมฟอเรส (Random Forest) ที่สร้างขึ้นจากลายพิมพ์ระดับโมเลกุลทั้งสองรูปแบบคือลายพิมพ์ระดับโมเลกุลแบบ Pubchem และแบบ Substructure

ทดสอบแบบจำลองแล้วบันทึกผลเป็น confusion matrix



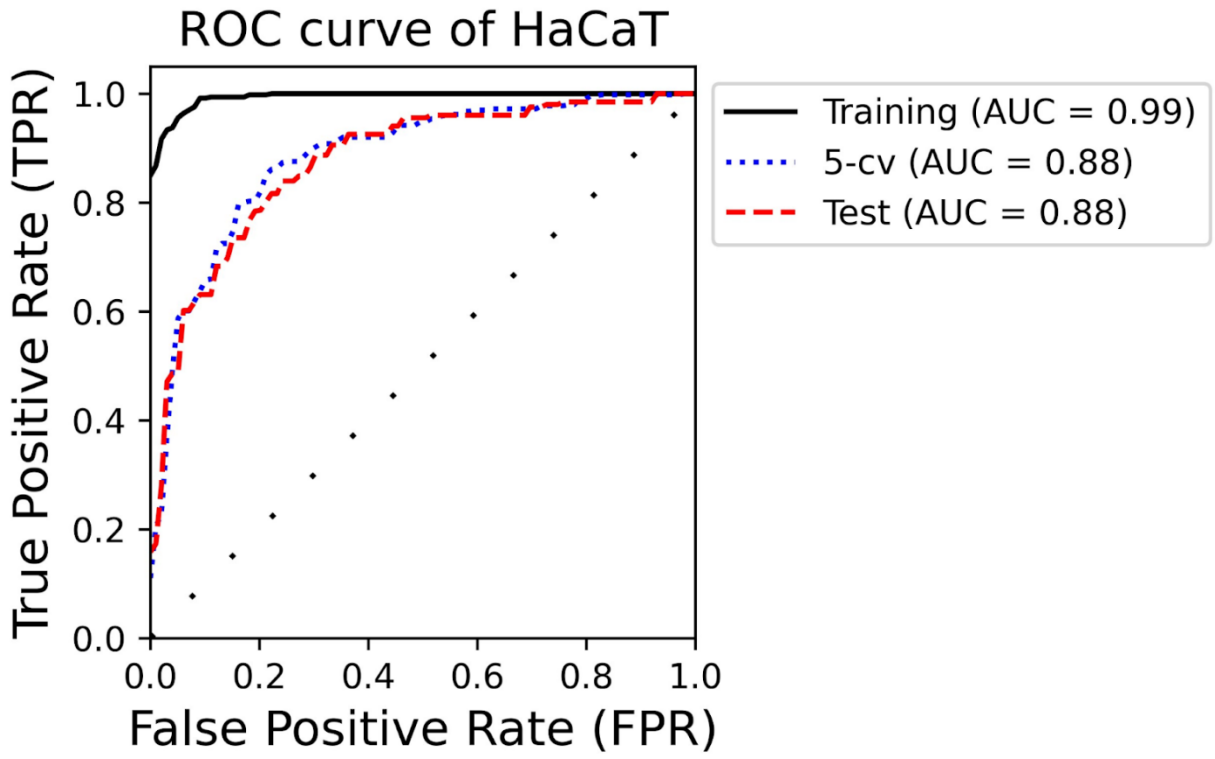
รูปที่ 6 Confusion matrix ของแบบจำลองแรนดอมฟอเรส (Random Forest) ที่สร้างขึ้นจากลายพิมพ์ระดับโมเลกุลแบบ Pubchem



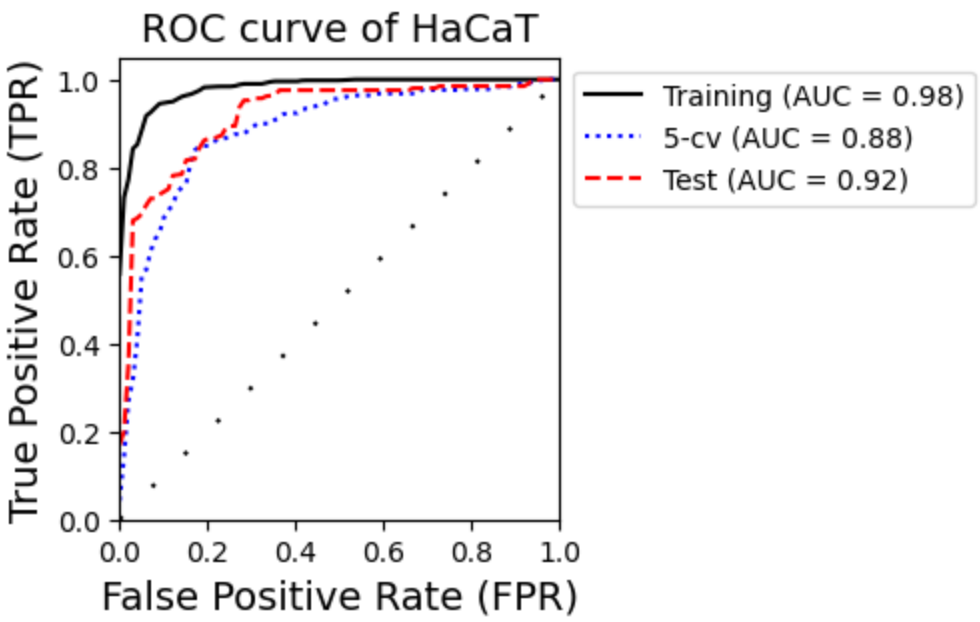
รูปที่ 7 Confusion matrix ของแบบจำลองแรนดอมฟอเรส (Random Forest) ที่สร้างขึ้นจากลายพิมพ์ระดับโมเลกุลแบบ Substructure

ใช้อัลกอลิทึมสร้าง Receiver Operating Characteristics (ROC) Curve ของแบบจำลองแรนดอมฟอเรส (Random Forest) ที่สร้างขึ้นจากลายพิมพ์ระดับโมเลกุลทั้งสองรูปแบบคือลายพิมพ์ระดับโมเลกุลแบบ Pubchem และแบบ Substructure ด้วยภาษา Python ดังรูปที่ 7 และ 8 โดยผลลัพธ์ทั้งสองบ่งชี้ให้เห็นว่า

แบบจำลองทั้งสองที่สร้างขึ้นมีความสามารถในการทำนายได้ดีเนื่องจากมีค่า Area under the curve (AUC) มากกว่าร้อยละ 88



รูปที่ 7 Receiver Operating Characteristics (ROC) Curve ของแบบจำลองแรนดอมฟอเรส (Random Forest) ที่สร้างขึ้นจากลายพิมพ์ระดับโมเลกุลแบบ Pubchem



รูปที่ 8 Receiver Operating Characteristics (ROC) Curve ของแบบจำลองแรนดอมฟอเรส (Random Forest) ที่สร้างขึ้นจากลายพิมพ์ระดับโมเลกุลแบบ Substructure