Rattanakosin Journal of Science and Technology: RJST
2025, Volume 7 Issue 1: 78-88
Research Article
Online ISSN: 2673-0553

การศึกษาเปรียบเทียบประสิทธิภาพตัวแบบการเรียนรู้ของเครื่องในการจำแนกผู้ป่วยโรคอัลไซเมอร์

A Comparison Study of the Performance of Machine Learning Models for Alzheimer's Disease Classification

กานต์ณัฐ ณ บางช้าง[1*]

Kannat Na Bangchang[1*]

[1]สาขาวิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์

[1]Department of Mathematics and Statistics, Faculty of Science and Technology
Thammasat University, Rangsit

[*]E-mail: kannat@mathstat.sci.tu.ac.th

บทคัดย่อ

งานวิจัยนี้ศึกษาและเปรียบเทียบประสิทธิภาพตัวแบบการเรียนรู้ของเครื่องในการจำแนกผู้ป่วยโรคอัลไซเมอร์ เนื่องจากโรคอัลไซเมอร์เป็นสาเหตุที่พบได้บ่อยที่สุดของภาวะสมองเสื่อม ซึ่งส่งผลให้ความสามารถในการใช้ชีวิตประจำวันของผู้ป่วยถดถอยลงเป็นอย่างมาก โดยจาก 55 ล้านคนทั่วโลกที่เป็นภาวะสมองเสื่อม 60-70% มีสาเหตุมาจากโรคอัลไซเมอร์ ผู้วิจัยจึงต้องการจำแนกผู้ป่วยมารักษาให้ได้อย่างทันท่วงที โดยจะทดลองใช้ชุดข้อมูล Darwin ที่มีตัวแปรอิสระทั้งหมด 451 ตัวแปร มีขนาดตัวอย่างทั้งหมด 174 คน และใช้โปรแกรมไพธอน (Python) ในการวิเคราะห์ข้อมูล โดยจากการสำรวจข้อมูล พบว่า ข้อมูลมีปัญหา High Dimensional และมีปัญหา Multicollinearity จึงได้ใช้เทคนิคการวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis) ในการแก้ไขปัญหาเหล่านี้ จากนั้นใช้เทคนิคการเรียนรู้ของเครื่องในการทำนายว่าข้อมูลตัวอย่างเป็นโรคอัลไซเมอร์หรือไม่ โดยทำการเปรียบเทียบในตัวแบบป่าสุ่ม (Random Forest) การถดถอยโลจิสติก (Logistic Regression) และ XGBoost ผลการวิจัย พบว่า ตัวแบบ Logistic Regression มีประสิทธิภาพดีที่สุดในการจำแนกผู้ป่วยโรคอัลไซเมอร์ในการวิเคราะห์ชุดข้อมูลทดสอบ โดยให้ค่า Accuracy Precision Recall และ F1-score เป็น 88.57%, 100%, 80.00% และ 88.89% ตามลำดับ

คำสำคัญ: ตัวแบบการเรียนรู้ของเครื่อง โรคอัลไซเมอร์ การจำแนก

* Corresponding author, e-mail: kannat@mathstat.sci.tu.ac.th

Received 21 06 2024 / Revised: 30 11 2024 / Accepted: 09 01 2025

Abstract

This research studied and compared the performance of machine learning models for Alzheimer's disease classification. Alzheimer 's is the most founded in the disease of the brain. Moreover, it affects the daily life of people. From much research indicated that there are about 55 million people in the world that have Dementia disease, 60 to 70 percent of them are caused Alzheimer's disease. Researcher studied the classification for urgent on the way of treatment. Darwin dataset is used to compare in this study. There are 451 covariates and 174 observations. Python is used in this study. Moreover, those covariates are multicollinearity and high dimensional dataset. The Principal Component Analysis is the technique for dealing first and then the supervise machine learning for classification on the outcome of Alzheimer. Those methods to compare contain Random Forest, logistic regression and XGBoost. The results pointed that logistic regression yields the high performance for classification. The Accuracy is 88.57%, precision is 100%, recall is 80.00% and F1-score is 88.89%, respectively.

Keywords: Machine learning model, Alzheimer 's disease, Classification

1. บทนำ

โรคอัลไซเมอร์เป็นโรคทางสมองที่จะมีอาการรุนแรงขึ้นเมื่อเวลาผ่านไป เนื่องจากโรคอัลไซเมอร์ทำให้สมองหดตัวและเซลล์สมองตายในที่สุด โรคอัลไซเมอร์เป็นสาเหตุที่พบบ่อยที่สุดของภาวะสมองเสื่อม โดยความจำ การคิด พฤติกรรม และทักษะทางสังคมจะค่อย ๆ ลดลง การเปลี่ยนแปลงเหล่านี้ส่งผลต่อความสามารถในการทำงานของผู้ป่วย [1] โดยประมาณ 6.5 ล้านคนในสหรัฐอเมริกาที่มีอายุ 65 ปีขึ้นไปเป็นโรคอัลไซเมอร์ ซึ่งในจำนวนนี้มากกว่าร้อยละ 70 มีอายุ 75 ปีขึ้นไป จากจำนวนผู้ป่วยโรคสมองเสื่อมทั่วโลกประมาณ 55 ล้านคน คาดว่าร้อยละ 60-70 เป็นโรคอัลไซเมอร์ [2] จากข้อมูลจำนวนผู้ป่วยโรคอัลไซเมอร์และอาการของโรคข้างต้น ทำให้ผู้วิจัยมีความสนใจที่จะจำแนกผู้ป่วยโรคอัลไซเมอร์มารักษาให้ได้อย่างทันท่วงที โดยผู้วิจัยต้องการหาวิธีการวิเคราะห์ข้อมูลได้อย่างรวดเร็วและถูกต้อง โดยการวิเคราะห์ข้อมูลเพื่อจำแนกผู้ป่วยโรคอัลไซเมอร์ว่าเป็นโรคอัลไซเมอร์หรือไม่นั้น จะทดลองวิเคราะห์ชุดข้อมูลของผู้ป่วยโรคอัลไซเมอร์ที่ชื่อว่า Darwin ซึ่งมีข้อมูลที่มีจำนวนของตัวแปรอิสระเป็นจำนวนมาก ในงานวิจัยนี้จึงใช้โปรแกรมไพธอน (Python) ในการวิเคราะห์ข้อมูลและใช้เทคนิคการแปลงคุณลักษณะในการเตรียมข้อมูลให้สามารถวิเคราะห์ได้ดียิ่งขึ้น โดยเทคนิควิธีที่ใช้ในการวิเคราะห์เบื้องต้น คือ วิธีการวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis: PCA) ซึ่งเป็นเทคนิคที่ได้รับความนิยมในการจัดการคุณลักษณะให้ได้ชุดข้อมูลที่เน้นไปที่คุณลักษณะที่มีความสัมพันธ์ในการวิเคราะห์ที่สำคัญที่สุด [3] หลังจากเตรียมชุดข้อมูลจะนำชุดข้อมูลที่ได้มาทดสอบตัวแบบด้วยเทคนิคการเรียนรู้ของเครื่อง (Machine Learning) โดยเทคนิคการเรียนรู้ของเครื่องเป็นวิวัฒนาการก้าวกระโดดทางเทคโนโลยีที่รวดเร็วและมีศักยภาพ ทำให้ได้รับความสนใจอย่างมากในช่วง 10 ปีนี้ [4] เนื่องจากเทคนิคการเรียนรู้ของเครื่องเป็นการประยุกต์ใช้สถิติขั้นสูงเพื่อเรียนรู้ให้สามารถระบุรูปแบบและคาดการณ์

ได้อย่างยอดเยี่ยม โดยเทคนิคการเรียนรู้ของเครื่องเป็นระบบที่สามารถเรียนรู้ได้จากตัวอย่างด้วยตนเองปราศจากการป้อนคำสั่งของผู้เขียนโปรแกรม ซึ่งมีความแตกต่างจากการเขียนโปรแกรมสมัยก่อน การเขียนโปรแกรมสมัยก่อนนั้นแนวทางทั้งหมดจะต้องถูกกำหนดไว้ชัดเจนด้วยผู้เชี่ยวชาญเพียงคนเดียว แต่จากอุตสาหกรรมซอฟต์แวร์ที่ถูกพัฒนาขึ้นจะมีการนำพื้นฐานความเข้าใจด้านตรรกศาสตร์ โดยโปรแกรมจะทำงานและส่งผลลัพธ์ออกมาตามคำสั่งที่สอดคล้องกับตรรกะ นอกจากนี้เมื่อระบบมีความซับซ้อนมากขึ้นก็จะมีความยุ่งยากในการเขียนโปรแกรม ดังนั้นจึงสามารถสรุปได้ว่าเทคนิคการเรียนรู้เครื่องมีประสิทธิภาพมากกว่าวิธีการเขียนโปรแกรมแบบดั้งเดิม โดยโปรแกรมจะเรียนรู้ข้อมูลทั้งหมดที่มีความเกี่ยวข้องกันและสร้างแนวทางที่มีความเหมาะสมกับข้อมูลได้โดยที่ผู้เขียนโปรแกรมไม่จำเป็นต้องเขียนโปรแกรมใหม่ทุกครั้งที่มีข้อมูลใหม่ โดยโปรแกรมจะปรับตัวเข้ากับข้อมูลใหม่ ในงานวิจัยครั้งนี้ผู้วิจัยใช้การวิเคราะห์ด้วยการใช้เทคนิคการจำแนกประเภทโดยการเรียนรู้ของเครื่อง ได้แก่ ป่าสุ่ม (Random Forest) การถดถอยโลจิสติก (Logistic Regression) และ XGBoost โดยทำการเปรียบเทียบประสิทธิภาพของตัวแบบด้วยเกณฑ์ในการพิจารณา Accuracy Precision Recall และ F1-score

2. ทฤษฎีที่เกี่ยวข้องในงานวิจัย

2.1 ป่าสุ่ม (Random Forest) คือ การสร้างตัวแบบที่ประกอบด้วยหลาย ๆ ต้นไม้ตัดสินใจ (Decision Tree) โดยแต่ละ Decision Tree จะถูกสร้างด้วยข้อมูลสุ่ม (random sample) และ attribute สุ่ม (random subset) ทำให้ไม่มี Decision Tree ใดที่เหมือนกันอย่างแน่นอน [5] การสร้างตัวแบบป่าสุ่ม จะเริ่มต้นด้วยการสุ่มข้อมูล (bootstrap sampling) เพื่อสร้างชุดข้อมูลสำหรับแต่ละ Decision Tree ที่ไม่เหมือนกัน จากนั้นใช้เทคนิค Bagging (Bootstrap Aggregating) ในการสร้าง Decision Tree แต่ละต้น โดยการสุ่มเลือก attribute มาแบ่งข้อมูลในแต่ละ node ในการใช้งาน Random Forest สามารถนำไปใช้งานได้หลากหลาย เช่น การจำแนกและทำนายผลของข้อมูล (classification and prediction) หรือใช้ในการค้นหาความสัมพันธ์ของ attribute กับผลลัพธ์ โดยที่แต่ละ Decision Tree ใน Random Forest จะเห็นภาพรวมของข้อมูลที่แตกต่างกัน ทำให้สามารถลดความผิดพลาดของตัวแบบได้มากขึ้น นอกจากนี้เทคนิคป่าสุ่ม ยังสามารถทำ Feature Importance ได้ คือ การคำนวณความสำคัญของแต่ละ attribute โดยใช้ Gini importance หรือ Mean Decrease Impurity (MDI) โดยที่ค่าสำคัญของแต่ละ attribute จะบ่งบอกถึงความสำคัญของ attribute ในการแบ่งข้อมูลในแต่ละ node ของ Decision Tree แต่ละต้นในตัวแบบป่าสุ่ม

2.2 การถดถอยโลจิสติก (Logistic Regression) เป็นเทคนิคทางสถิติเพื่อการจำแนกชุดข้อมูลเป็นสองกลุ่ม โดยที่ Logistic regression เป็นการตั้งชื่อตามคำสั่งที่ชื่อว่า Logistic function นอกจากนี้ยังถูกเรียกว่า Sigmoid function ที่ถูกพัฒนาโดยนักสถิติเพื่อบรรยายคุณลักษณะของการเติบโตของประชากรในทางเศรษฐศาสตร์ที่มีการเพิ่มขึ้นอย่างรวดเร็วและถึงจุดสูงสุดที่ทำตามความสามารถของทรัพยากรที่มี ซึ่งทำให้เกิดแผนภูมิแบบเส้นรูปร่างคล้ายตัวอักษรเอสในภาษาอังกฤษ (S-shaped curve) ที่สามารถแสดงถึงตัวเลขที่เป็นค่าที่อยู่ระหว่าง 0 และ 1 [6] แสดงได้ดังนี้

$$\frac{1}{1+e^{-value}}$$

เมื่อ e คือ ฐานของลอการิทึมธรรมชาติ(natural logarithms) และ value คือ ตัวเลขที่ต้องการแปลง

การถดถอยโลจิสติก (Logistic regression) ใช้แนวคิดที่มีความคล้ายคลึงกับการถดถอยเชิงเส้นตรง (linear regression) มาก โดยที่ตัวแปรอิสระ (x) จะถูกรวมกันเพื่อใช้ในการให้น้ำหนักและประเมินความสัมพันธ์กับตัวแปรตาม (y) ข้อแตกต่างที่สำคัญจาก linear regression คือ ค่าตัวแปรตามจะถูกเปลี่ยนให้เป็นค่าสองค่า คือ 0 หรือ 1 โดยตัวอย่างสมการ logistic regression เป็นดังนี้

$$y = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

เมื่อ y คือ ค่าตัวแปรตามจากการทำนาย

β_0 คือ ค่า bias หรือ intercept

β_1 คือ ค่าสัมประสิทธิ์ (coefficient) สำหรับตัวแปรอิสระตัวที่ 1 (X)

โดยค่าสัมประสิทธิ์จะได้จากการเรียนรู้ของเครื่องจากชุดข้อมูลเรียนรู้

2.3 XGBoost (eXtreme Gradient Boosting) คือ กระบวนการเรียนรู้ของเครื่องที่ได้รับความนิยมสูงสุดในปัจจุบัน โดยเป็นเครื่องมือที่ใช้ในการทำนายแบบ Supervised Learning ซึ่งสามารถใช้กับปัญหาการจำแนกหรือการทำนายที่มีข้อมูลขนาดใหญ่และมีความซับซ้อน มักนิยมใช้ในการแข่งขันด้านการทำนายเชิงแข่งขัน [7] XGBoost เป็นเทคนิค Ensemble Learning ที่ประกอบไปด้วย Decision Trees หลายต้นที่มีการใช้ Gradient Boosting Algorithm ในการเพิ่มความแม่นยำของ Decision Trees แต่ละต้น ซึ่ง XGBoost มีความสามารถในการทำ Feature Selection และ Regularization ที่ช่วยลด Overfitting และเพิ่มประสิทธิภาพของตัวแบบได้มากขึ้น XGBoost ใช้ Gradient Boosting Algorithm ในการเพิ่มความแม่นยำของ Decision Trees แต่ละต้น โดยใช้วิธีการนำค่าคงที่ learning rate เข้ามาช่วยปรับค่าความคลาดเคลื่อนในแต่ละรอบการสร้างตัวแบบ (Train Model) ซึ่งจะทำให้ตัวแบบมีการเรียนรู้ที่เร็วขึ้น และลดความผิดพลาดในการทำนายได้ นอกจากนี้ XGBoost ยังสามารถใช้ Gradient-based One-Side Sampling (GOSS) และ Weight-based Random Sampling (WBRS) เพื่อช่วยลดเวลาในการทดสอบตัวแบบ และป้องกัน Overfitting โดยทำการลดจำนวนข้อมูลที่ใช้ทดสอบ และยังสามารถใช้ Early Stopping เพื่อหยุดการทดสอบตัวแบบที่มีค่าคลาดเคลื่อนเพิ่มเติมได้

2.4 การวัดประสิทธิภาพของตัวแบบการพยากรณ์

เกณฑ์วัดประสิทธิภาพที่ใช้ในการวัดประสิทธิภาพของตัวแบบการพยากรณ์แบบแบ่งประเภท (classification) ในงานวิจัยนี้มีอยู่ 4 ค่า โดยอ้างอิงจาก [8] ดังนี้

2.4.1 Recall คือ การวัดความถูกต้องของตัวแบบพยากรณ์ แล้วคำนวณจากการที่ตัวแบบพยากรณ์สามารถตอบคำถามถูกในส่วนที่สนใจได้มากน้อยเพียงใดจากค่าที่ถูกต้องจริงทั้งหมด สามารถคำนวณได้จาก Recall = True Positive / (True Positive + False Negative)

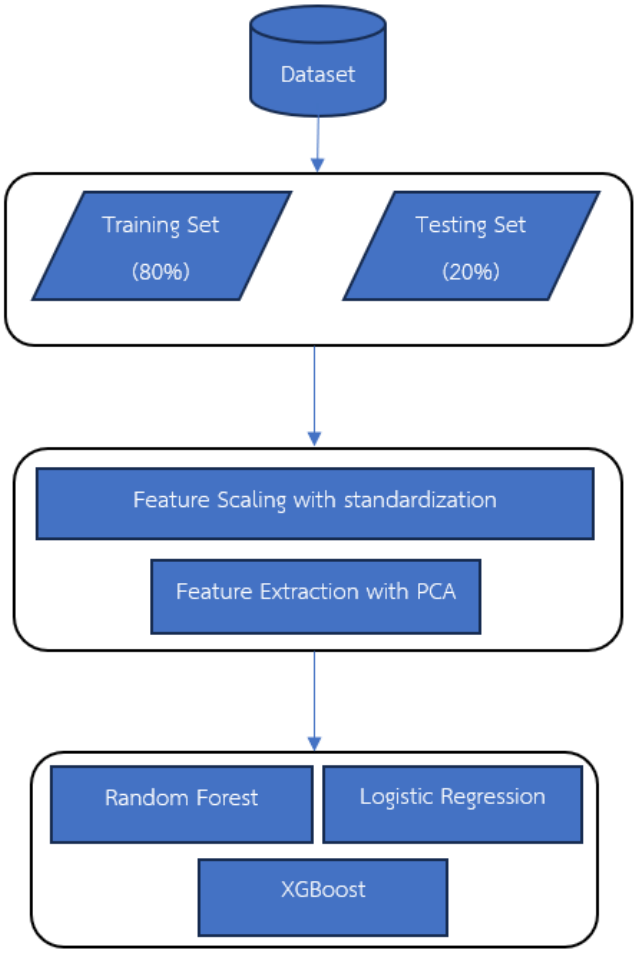
2.4.2 Precision คือ การวัดความแม่นยำของตัวแบบพยากรณ์ แล้วคำนวณจากการที่ตัวแบบพยากรณ์สามารถตอบคำถามถูกได้ถูกในส่วนที่ไม่ได้สนใจมากน้อยเพียงใดจากค่าที่ถูกต้องจริงทั้งหมด สามารถคำนวณได้จาก Precision = True Positive / (True Positive + False Positive)

2.4.3 F1-score คือ การวัดค่า Precision และ Recall พร้อมกันของตัวแบบพยากรณ์ โดยที่ทำการพิจารณาแยกทีละผลลัพธ์ที่สนใจและไม่สนใจ ซึ่งใช้เป็นตัวบอกถึงความเหมาะสมของการนำข้อมูลไปใช้ว่ามีความเหมาะสมเพียงใด พร้อมทั้งบอกถึงประสิทธิภาพของตัวแบบพยากรณ์ [9] สามารถคำนวณได้จาก F1-score = (2 x Precision x Recall) / (Precision + Recall)

2.4.4 Accuracy คือ การวัดความถูกต้องของตัวแบบพยากรณ์ โดยพิจารณารวมทุกผลลัพธ์ที่เกิดขึ้นทั้งหมด เพื่อดูความแม่นยำในการทำนายผลถูกต้องมากน้อยเพียงใด [10] สามารถคำนวณได้จาก Accuracy = (True Positive + Negative Positive) / จำนวนชุดข้อมูลทั้งหมด

3. วิธีดำเนินการวิจัย

วิธีดำเนินการวิจัยในการศึกษานี้ ประกอบด้วย การทำความเข้าใจข้อมูล การเตรียมข้อมูล และการสร้างแบบจำลองข้อมูล ซึ่งขั้นตอนในการวิจัยทั้งหมด สรุปได้ดังรูปที่ 1



รูปที่ 1 แสดงขั้นตอนการทำงาน

3.1 การทำความเข้าใจข้อมูล (Data Understanding) ประกอบด้วย การอธิบายชุดข้อมูล และการสำรวจข้อมูล

3.1.1 การอธิบายชุดข้อมูล (Data Description) ในส่วนของการอธิบายชุดข้อมูล ข้อมูลที่เราใช้ในงานนี้ได้รับการจัดเรียงตามระดับความยาก โดยคำนึงถึงความต้องการในการใช้ความคิดเชิงประสาทของผู้รับการทดสอบ มีการแบ่งงานออกเป็นกลุ่มต่าง ๆ ตามวัตถุประสงค์ ดังนี้

1. งานทางกราฟฟิก ทดสอบความสามารถของผู้เข้าร่วมในการเขียนระดับประถมศึกษา รวมถึงการเชื่อมจุดต่าง ๆ และการวาดรูปทรงเรขาคณิต

2. งานคัดลอก ประเมินความสามารถของผู้เข้าร่วมในการทำซ้ำที่ซับซ้อน เช่น ตัวอักษร คำ และตัวเลข

3. งานทดสอบความจำ ทดสอบการเปลี่ยนแปลงในกระบวนการเขียนก่อนหน้านี้ จดจำวัตถุที่แสดงในภาพ

4. งานเขียนตามคำบอก ตรวจสอบว่าการเขียนด้วยลายมือแตกต่างกันอย่างไรเมื่อใช้หน่วยความจำในการทำงาน

ชุดข้อมูลประกอบด้วยข้อมูลจากผู้รับการทดสอบทั้งหมด 174 คน โดย 89 คนเป็นผู้ป่วยโรคอัลไซเมอร์ (AD) และ 85 คน เป็นคนปกติที่ไม่มีโรคอัลไซเมอร์ 25 ข้อ โดยเป็นข้อมูลที่ศึกษาเกี่ยวกับตัวแปรที่เกี่ยวข้องกับการใช้ทักษะในด้านต่าง ๆ เพื่อที่จะตรวจสอบว่าตัวแปรใดบ้างที่มีความแตกต่างกันในกลุ่มผู้ป่วยอัลไซเมอร์และในคนปกติ โดยมีเกณฑ์ต่าง ๆ ที่ใช้ในการประเมินในแต่ละทักษะ ซึ่งชุดข้อมูลนี้เป็นตัวอย่างที่ชี้ให้เห็นถึงกรณีข้อมูลที่มีมิติสูง กล่าวคือ มีจำนวนตัวแปรอิสระทั้งหมด 451 ตัวแปร และมีขนาดตัวอย่างเท่ากับ 174 คน [10]

3.1.2 การสำรวจข้อมูล (Data Exploration) การสำรวจข้อมูลในขั้นตอนนี้เพื่อความเข้าใจในชุดข้อมูลเบื้องต้น และสำรวจเพื่อใช้ในการเตรียมข้อมูลในขั้นตอนต่อไป การสำรวจข้อมูลในเบื้องต้นเพื่อความเข้าใจในทั้งความหมายและลักษณะของข้อมูลในแต่ละคุณลักษณะ รวมถึงค่าต่าง ๆ ในคุณลักษณะด้วย

1. การสำรวจเพื่อค้นหาและการแก้ไขค่าว่าง (Missing Value)

2. การสำรวจการจำแนกกลุ่มของข้อมูล (Classification)

3. การสำรวจเพื่อค้นหาปัญหาของตัวแปร

3.2 การเตรียมข้อมูล (Data Preparation) เป็นการเตรียมข้อมูลให้เหมาะสมก่อนการสร้างแบบจำลองข้อมูลด้วยการเรียนรู้ของเครื่อง ประกอบด้วย

1. การลบคอลัมน์ที่ไม่ได้ใช้ในการวิเคราะห์ คือ นำคอลัมน์ ID ออก

2. การแปลงข้อมูล (Data Transformation) แปลงข้อมูลในคอลัมน์ class ให้เป็นตัวเลข โดยที่ Patient เป็น 1 และ Healthy เป็น 0

3. การแบ่งชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ (Data Splitting) เพื่อนำชุดข้อมูลเรียนรู้เข้าสู่กระบวนการเรียนรู้ต่าง ๆ ของเครื่องในขั้นตอนต่อไป ส่วนชุดข้อมูลทดสอบมีไว้เพื่อทดสอบกับแบบจำลองข้อมูลต่าง ๆ ที่สร้างขึ้นในขั้นตอนสุดท้าย ทั้งนี้ กำหนดให้การแบ่งชุดข้อมูลเป็นแบบสุ่ม

4. การสกัดคุณลักษณะของข้อมูล (Feature Extraction)

ในการศึกษานี้จะทำการสกัดคุณลักษณะโดยวิธีการวิเคราะห์ตัวประกอบหลัก (PCA) เนื่องจากเทคนิคนี้จะช่วยในการลดจำนวนตัวแปรอิสระ แต่ตัวแปรที่เหลือยังสามารถอธิบายตัวแปรตามได้สูง ซึ่งนอกจากช่วยลดความซับซ้อนก่อนนำไปใช้กับวิธีการทาง Machine Learning แล้วยังช่วยในการลดความคลาดเคลื่อนของการนำตัวแบบไปใช้ในกรณีที่จำนวนตัวแปรอิสระมากเกินไป โดยมีขั้นตอนดังนี้

1. การปรับช่วงขอบเขตของข้อมูลชนิดตัวเลข (Feature Scaling) เพื่อปรับข้อมูลตัวเลขให้อยู่ในช่วงเดียวกัน ทำให้เมื่อนำเข้าข้อมูลเข้าสู่การเรียนรู้ของเครื่องแล้ว ทุกคุณลักษณะจะมีน้ำหนักเท่าเทียมกัน โดยเฉพาะอย่างยิ่งในแบบจำลองข้อมูลที่เลือกใช้วิธีการจำแนกด้วย logistic regression เนื่องจากข้อมูลมีการแจกแจงแบบไม่ปกติ (Non-normal distribution) จึงเลือกการปรับช่วงขอบเขตของข้อมูลด้วยการใช้วิธีการ Standardize ซึ่งเป็นวิธีการปรับขนาดตามค่าเฉลี่ยและ ค่าส่วนเบี่ยงเบนมาตรฐาน

2. หาเมทริกซ์ความแปรปรวนร่วม (Covariance matrix)

3. หาไอเกนแวลู (Eigen Value) และไอเกนเวกเตอร์ (Eigen Vector) จากเมทริกซ์ความแปรปรวนร่วม

4. เลือกจำนวนตัวประกอบที่เหมาะสม กำหนดจำนวนของส่วนประกอบที่เหมาะสม พิจารณาความแปรปรวนที่ได้รับการอธิบายทั้งหมด การเลือกที่นิยม คือ การเลือกจำนวนของส่วนประกอบที่ความแปรปรวนที่ได้รับการอธิบายมีค่ามากกว่าหรือเท่ากับ 80%

5. Project Data โดยการเรียกใช้ pca library ของแพกเกจ scikit-learn หรือจะทำการคำนวณมือโดยใช้สมการ $\mathrm{Pr}ojectData = X\_train\_scaled.dot(principle\_components)$ ในโปรแกรมภาษาไพธอน

เมื่อ $X\_train\_scaled.dot$ คือ ข้อมูลชุดทดสอบที่ทำการปรับช่วงข้อมูลแล้ว และ $principle\_components$ คือ ไอเกนเวกเตอร์จากการเลือกจำนวนตัวประกอบ

3.3 การสร้างแบบจำลองข้อมูล (Modeling) การศึกษานี้เลือกแบบจำลองข้อมูลในการเปรียบเทียบสมรรถนะของแบบจำลองข้อมูลใน 3 ตัวแบบ ดังนี้ Random Forest Logistic Regression และ XGBoost แบ่งออกเป็นการวัดผลในชุดข้อมูลเรียนรู้และในชุดข้อมูลทดสอบ โดยผู้วิจัยเลือกใช้ 3 ตัวแบบดังกล่าว โดยพิจารณาเลือกตัวแบบพื้นฐานที่ใช้ในการพยากรณ์ในตัวแปรตามที่เป็นแบบ Binary คือ Logistic Regression และพิจารณาเลือกตัวแบบที่ใช้วิธีการทางด้าน Machine Learning จำนวน 2 ตัวแบบ คือ Random Forest และ XGBoost โดยไม่ได้นำตัวแบบอื่น ๆ เช่น Support Vector Machines มาใช้เนื่องจากจะต้องมีการกำหนด hyperparameter ซึ่งจำเป็นที่จะต้องกำหนดให้สอดคล้องกับคุณสมบัติของตัวแปรที่เกี่ยวข้องทั้งหมด

3.3.1 การวัดผลในชุดข้อมูลเรียนรู้ นำชุดข้อมูลเรียนรู้มาทดสอบประสิทธิภาพการจำแนกด้วยการแบ่งสุ่มข้อมูลเพื่อเรียนรู้และทดสอบในชุดข้อมูลเรียนรู้ ด้วยการใช้ 10-fold cross-validation ในแต่ละแบบจำลองข้อมูล โดยวัดผลการจำแนกด้วยค่าเฉลี่ยของ Accuracy Precision Recall และ F1-score ของการทดสอบทั้งหมดแล้วนำค่าเฉลี่ยดังกล่าวเพื่อใช้ในการเปรียบเทียบประสิทธิภาพการจำแนกของแต่ละแบบจำลองข้อมูล

3.3.2 การวัดผลในชุดข้อมูลทดสอบ เมื่อสร้างแบบจำลองจากชุดข้อมูลเรียนรู้แล้ว จะนำแบบจำลองที่ได้มาทดสอบกับชุดข้อมูลทดสอบที่ได้แยกไว้ก่อนการเรียนรู้ เพื่อวัดผลจำแนกด้วยค่า Accuracy Precision Recall และ F1-score

4. ผลการวิจัย

แบ่งออกเป็น 2 ส่วน คือ ผลการประเมินประสิทธิภาพในแต่ละตัวแบบในข้อมูลชุดเรียนรู้และข้อมูลชุดทดสอบ

ตารางที่ 1 ผลการประเมินประสิทธิภาพในแต่ละตัวแบบในข้อมูลชุดเรียนรู้

ตัวแบบ	Accuracy	Precision	Recall	F1-score
Random Forest	0.8416*	0.8613*	0.8283*	0.8350*
Logistic Regression	0.8132	0.8452	0.7690	0.7990
XGBoost	0.8203	0.8589	0.7833	0.8126

ตารางที่ 2 ผลการประเมินประสิทธิภาพในแต่ละตัวแบบในข้อมูลชุดทดสอบ

ตัวแบบ	Accuracy	Precision	Recall	F1-score
Random Forest	0.7429	1.0000*	0.5500	0.7097
Logistic Regression	0.8857*	1.0000*	0.8000*	0.8889*
XGBoost	0.8571	0.9412	0.8000*	0.8649

ตารางที่ 3 แสดง Confusion Matrix ของตัวแบบ Random Forest

Random Forest		Predict	
		Positive (Patient)	Healthy
Actual	Positive (Patient)	15	0
	Negative (Healthy)	9	11

ตารางที่ 4 แสดง Confusion Matrix ของตัวแบบ Logistic Regression

Logistic Regression		Predict	
		Positive (Patient)	Healthy
Actual	Positive (Patient)	14	1
	Negative (Healthy)	4	16

ตารางที่ 5 แสดง Confusion Matrix ของตัวแบบ XGBoost

XGBoost		Predict	
		Negative (Healthy)	Negative (Healthy)
Actual	Negative (Healthy)	14	1
	Negative (Healthy)	4	16

จากตารางที่ 1 ผลลัพธ์ที่ได้ของแต่ละตัวแบบสำหรับชุดข้อมูลชุดเรียนรู้ ด้วยการตรวจสอบโดยการ cross-validation พบว่า ตัวแบบ Random Forest นั้นให้ค่าวัดประสิทธิภาพสูงที่สุดในทั้ง 4 เกณฑ์ กล่าวคือ ให้ค่า Accuracy เท่ากับ 0.8416 Precision เท่ากับ 0.8613 Recall เท่ากับ 0.8238 และให้ค่า F1-score เท่ากับ 0.8350 ตามลำดับ ซึ่งแสดงว่าตัวแบบ Random Forest นั้นให้ผลดีที่สุดจากทั้ง 3 ตัวแบบ ในทางกลับกันจะพบว่าตัวแบบ Logistic Regression นั้นให้ประสิทธิภาพต่ำที่สุดในทั้ง 4 เกณฑ์ คือ ให้ค่า Accuracy เท่ากับ 0.8132 Precision เท่ากับ 0.8452 Recall เท่ากับ 0.7690 และให้ค่า F1-score เท่ากับ 0.7990 นั่นแสดงว่าประสิทธิภาพของตัวแบบ Logistic Regression เป็นตัวแบบให้ผลการทำนายที่ต่ำสุดในข้อมูลชุดเรียนรู้

จากตารางที่ 2 ผลลัพธ์ที่ได้ของแต่ละตัวแบบในข้อมูลชุดทดสอบ พบว่าตัวแบบ Logistic Regression นั้นให้ค่าวัดประสิทธิภาพสูงที่สุดในทั้ง 4 เกณฑ์ ดังนี้ ให้ค่า Accuracy เท่ากับ 0.8857 Precision เท่ากับ 1.0000 Recall เท่ากับ 0.8000 และให้ค่า F1-score เท่ากับ 0.8889 ซึ่งแสดงว่าตัวแบบ Logistic Regression นั้นให้ผลการทำนายในชุดข้อมูลทดสอบ (unseen data) ดีที่สุดจากทั้ง 3 ตัวแบบ

จากตารางที่ 3 ถึง ตารางที่ 5 แสดง Confusion Matrix ของผลการนำแต่ละตัวแบบไปใช้ในการพยากรณ์ ซึ่งสอดคล้องกับผลลัพธ์ในการประเมินประสิทธิภาพของแต่ละตัวแบบ

5. สรุปผลการวิจัย

การศึกษานี้ได้ทำการวิเคราะห์และเปรียบเทียบประสิทธิภาพของตัวแบบการเรียนรู้เชิงกำหนดที่ใช้ในการจำแนกโรคอัลไซเมอร์ระหว่าง XGBoost Random Forest และ Logistic Regression โดยใช้ชุดข้อมูล Darwin ที่มีตัวแปรอิสระทั้งหมด 451 ตัวแปร มีจำนวนตัวอย่างทั้งหมด 174 คน โดยเป็นกรณีข้อมูลที่มีมิติสูง (High Dimensional data) ซึ่งในกรณีถ้าใช้การวิเคราะห์ข้อมูลด้วยเทคนิควิธีการดั้งเดิมจะส่งผลต่อค่าประมาณพารามิเตอร์ของตัวแบบจะคลาดเคลื่อนสูง และทำให้การนำตัวแบบไปใช้ในการพยากรณ์มีประสิทธิภาพลดลง ซึ่งถูกแบ่งเป็นสองส่วน คือ ข้อมูลชุดเรียนรู้และข้อมูลชุดทดสอบ ในอัตราส่วน 80:20 รวมทั้งการใช้ Cross Validation แบบ 10-fold เพื่อเพิ่มความน่าเชื่อถือของการทดสอบตัวแบบ จากผลการทดสอบแสดงให้เห็นว่าในข้อมูลชุดเรียนรู้ Random Forest จะเป็นตัวแบบที่มีประสิทธิภาพสูงที่สุด ในขณะที่ Logistic Regression เป็นตัวแบบที่มีประสิทธิภาพสูงที่สุดในข้อมูลชุดทดสอบ แสดงถึงความแม่นยำและความสามารถในการจำแนกที่เป็นเอกลักษณ์ของตัวแบบ Logistic Regression อย่างไรก็ตามการปรับใช้และนำไปใช้ในการตัดสินใจทางการแพทย์ในอนาคตอาจมีความสำคัญอย่างมากในประเด็นระยะเวลาในการฝึกฝนของแต่ละตัวแบบ จากผลการวิจัย พบว่า มีความใกล้เคียงกันมากสำหรับระยะเวลาในการฝึกฝนของทั้งสามตัวแบบที่ใช้ในการวิจัยครั้งนี้ โดยมีข้อสรุปจากงานวิจัยถึงการนำตัวแบบ Logistic Regression ไปประยุกต์ใช้กับข้อมูลที่มีมิติสูง (High Dimensional data) [12] นอกจากนี้ผลลัพธ์นี้ยังเป็นแนวทางสำคัญสำหรับงานวิจัยและพัฒนาตัวแบบที่มีประสิทธิภาพสูงในด้านการจำแนกผู้ป่วยโรคอัลไซเมอร์ในอนาคต อย่างไรก็ตามผลลัพธ์ในส่วนของข้อมูลชุดทดสอบ ผลลัพธ์ที่ได้อาจมีความคลาดเคลื่อน เนื่องจากชุดข้อมูล Darwin มีจำนวนตัวอย่างค่อนข้างน้อยเมื่อเทียบกับจำนวนลักษณะของข้อมูล ดังนั้นจึงอาจมีการพิจารณาด้วยชุดข้อมูลที่มีขนาดใหญ่ขึ้น เพื่อพิจารณาตัวแบบที่มีความเหมาะสมกับแต่ละลักษณะของข้อมูล

6. เอกสารอ้างอิง

[1] ลักขณา อินทร์กลับ. (2557). เข้าใจและเข้าถึงผู้ป่วยโรคอัลไซเมอร์. *วารสารมหาวิทยาลัยคริสเตียน.* 20(3), 439-446.

[2] Matthews, K. A., Xu, W., Gaglioti, A. H., Holt, J.B., Croft, J. B., Mack, D., & McGuire, L. C. (2019). Racial and ethnic estimates of Alzheimer's disease and related dementias in the United States (2015-2060) in adults aged≥ 65 years. *Alzheimer's & Dementia*. 15(1), 17-24. https://doi.org/10.1016/j.jalz.2018.06.3063

[3] GÜNDOĞDU, S. (2022). Hepatitis C Disease Detection Based on PCA–SVM Model. *Hittite Journal of Science and Engineering.* 9(2), 111-116. https://doi.org/10.17350/HJSE19030000261

[4] Mamdouh Farghaly, H., Shams, M. Y., & Abd El-Hafeez, T. (2023). Hepatitis C Virus prediction based on machine learning framework: a real-world case study in Egypt. *Knowledge and Information Systems*. 65, 2595-2617. 10.1007/s10115-023-01851-4

[5] Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC medical informatics and decision making*. 19(1), 1-16. 10.1186/s12911-019-1004-8

[6] Nagavelli, U., Samanta, D., & Chakraborty, P. (2022). Machine learning technology-based heart disease detection models. *Journal of Healthcare Engineering*. 1-9. https://doi.org/10.1155/2022/7351061

[7] Meriç, E., & Özer, Ç. (2022). *Symptom Based Health Status Prediction via Decision Tree, KNN, XGBoost, LDA, SVM, and Random Forest*. Proceedings of International Conference on Computing Intelligence and Data Analytics, Koceli, Turkey, 193-207.

[8] Powers, D. M. (2008). Evaluation From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. *Internation Journal of Machine Learning Technology*. 2(1), 37-63. https://doi.org/10.48550/arXiv.2010.16061

[9] Asif, M. A. A. R., Nishat, M. M., Faisal, F., Dip, R. R., Udoy, M. H., Shikder, M. F., & Ahsan, R. (2021). Performance Evaluation and Comparative Analysis of Different Machine Learning Algorithms in Predicting Cardiovascular Disease. *Engineering Letters.* 29(2), 731-741.

[10] Syafa'ah, L., Zulfatman, Z., Pakaya, I., & Lestandy, M. (2021). Comparison of machine learning classification methods in hepatitis C virus. *Journal Online Informatika*. 6(1), 73-78. 10.15575/join.v6i1.719

[11] Cilia, N. D., De Gregorio, G., De Stefano, C., Fontanella, F., Marcelli, A., & Parziale, A. (2022). Diagnosing Alzheimer's disease from on-line handwriting: a novel dataset and performance benchmarking. *Engineering Applications of Artificial Intelligence*. 111, 1-12. https://doi.org/10.1016/j.engappai.2022.104822

[12] Mukherjee, S., Niu, Z., Halder, S. Bhattacharya, B. B., & Michailidis, G. (2024). *High Dimensional Logistic Regression Under Network Dependence*. https://doi.org/10.48550/arXiv.2110.03200