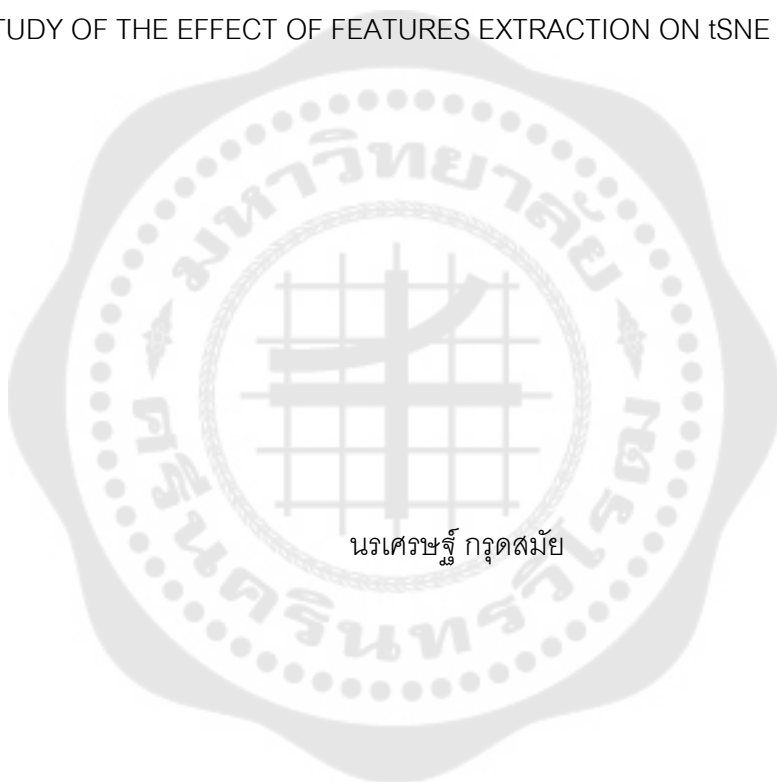




การศึกษาผลของการใช้อัลกอริทึมแยกแยะคุณลักษณะต่ออัลกอริทึม tSNE
STUDY OF THE EFFECT OF FEATURES EXTRACTION ON tSNE ALGORITHM



นรเศรษฐ์ กรุดสมัย

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

2563

การศึกษาลักษณะของการใช้อัลกอริทึมแยกแยะคุณลักษณะต่ออัลกอริทึม tSNE



สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2563
ลิขสิทธิ์ของมหาวิทยาลัยศรีนครินทรวิโรฒ

STUDY OF THE EFFECT OF FEATURES EXTRACTION ON tSNE ALGORITHM



NORASETH KRUDSAMAI

A Master's Project Submitted in Partial Fulfillment of the Requirements

for the Degree of MASTER OF SCIENCE

(Information Technology)

Faculty of Science, Srinakharinwirot University

2020

Copyright of Srinakharinwirot University

สารนิพนธ์

เรื่อง

การศึกษาผลของการใช้อัลกอริทึมแยกแยะคุณลักษณะต่ออัลกอริทึม tSNE

ของ

นรเศรษฐ์ กรุดสมัย

ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

ของมหาวิทยาลัยศรีนครินทรวิโรฒ

(รองศาสตราจารย์ นายแพทย์ฉัตรชัย เอกปัญญาสกุล)

คณบดีบัณฑิตวิทยาลัย

คณะกรรมการสอบปากเปล่าสารนิพนธ์

..... ที่ปรึกษาหลัก ประธาน
(อาจารย์ ดร.ศิริสรวพ เหล่าหะเกียรติ) (อาจารย์ ดร.สุทธิพงศ์ รัชชพงษ์)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.จันตรี ผลประเสริฐ)

ชื่อเรื่อง	การศึกษาผลของการใช้อัลกอริทึมแยกแยะคุณลักษณะต่ออัลกอริทึม tSNE
ผู้วิจัย	นรเศรษฐ์ กรุดสมัย
ปริญญา	วิทยาศาสตร์มหาบัณฑิต
ปีการศึกษา	2563
อาจารย์ที่ปรึกษา	อาจารย์ ดร. ศิริสรรพ เหล่าหะเกียรติ

ในปัจจุบันอัลกอริทึมสกัดคุณลักษณะ tSNE ได้รับความนิยมอย่างมากในการใช้ลดจำนวนมิติของชุดข้อมูลและการแสดงภาพของข้อมูล ทว่าการใช้อัลกอริทึม tSNE กับข้อมูลเพื่อลดจำนวนมิติลงก่อนทำการจำแนกข้อมูลจะทำให้ประสิทธิภาพการจำแนกข้อมูลลดลงอย่างมาก เนื่องจากอัลกอริทึม tSNE ไม่สามารถรักษาคุณสมบัติโดยรวมของข้อมูลไว้ได้ ในการศึกษาครั้งนี้เรานำเสนอการใช้อัลกอริทึมสกัดคุณลักษณะแบบสองชั้นเพื่อรักษาคุณสมบัติโดยรวมของข้อมูลไว้ อันจะเป็นการเพิ่มประสิทธิภาพในการจำแนกข้อมูลที่ผ่านการลดจำนวนมิติด้วยอัลกอริทึม tSNE จากการทดลองพบว่าผลการจำแนกข้อมูลด้วยการใช้อัลกอริทึมสกัดคุณลักษณะแบบสองชั้นให้ประสิทธิภาพดีกว่าการใช้อัลกอริทึม tSNE แต่เพียงอย่างเดียว อีกทั้งในบางชุดข้อมูลยังมีประสิทธิภาพดีกว่าการใช้ข้อมูลที่ไม่ผ่านการลดมิติอีกด้วย

คำสำคัญ : อัลกอริทึมสกัดคุณลักษณะ, การทำงานร่วมกันของอัลกอริทึมสกัดคุณลักษณะ, การลดมิติข้อมูล, การเรียนรู้ของเครื่อง

Title	STUDY OF THE EFFECT OF FEATURES EXTRACTION ON tSNE ALGORITHM
Author	NORASETH KRUDSAMAI
Degree	MASTER OF SCIENCE
Academic Year	2020
Thesis Advisor	Dr. Sirisup Laohakiat

Recently, the tSNE algorithm is one of the most popular feature extraction algorithms used in dimensionality reduction and data visualization tasks. However, when applying tSNE to reduce the dimensionality of the dataset before performing the classification, the classification results would substantially deteriorate. This is because tSNE cannot preserve the global structure of the dataset. In order to solve this drawback in tSNE, the use of a two-step feature extraction algorithm was proposed to improve the global structure of the reduced-dimension dataset. Another feature extraction algorithm was applied before performing tSNE in the final step. Then, the reduced dimension data is used to perform classification using various classification models, compared to using tSNE directly. The results showed that the classification results when using two-step feature extraction significantly improved in comparison to only using tSNE. Also, in some datasets, the classification results of using a two-step feature extraction on the full dimensional data.

Keyword : Feature Extraction Algorithms, Combined Feature Extraction Algorithms, Dimensionality Reduction, Machine Learning

กิตติกรรมประกาศ

สารนิพนธ์นี้สำเร็จลุล่วงได้ด้วยความช่วยเหลือจาก อ.ดร.ศิริสรรพ เหล่าหะเกียรติ อาจารย์ที่ปรึกษาที่ให้คำปรึกษา คำแนะนำในการทำสารนิพนธ์ ตลอดจนสนับสนุนข้อมูลทางวิชาการและข้อมูลสำหรับทำสารนิพนธ์นี้

ขอขอบคุณคณะกรรมการสอบสารนิพนธ์ ที่ได้ให้คำแนะนำและข้อเสนอแนะสำหรับการปรับปรุงสารนิพนธ์ และการใช้อัลกอริทึมแยกแยะคุณลักษณะ PCA, Kernel PCA, AUTOENCODER และ tSNE รวมไปถึงการใช้เทคนิคการเรียนรู้ของเครื่องโดยใช้แบบจำลอง Logistic Regression, Support Vector Machine, Decision Tree และ Random Forest

ขอขอบคุณบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒสำหรับทุนสนับสนุนการนำเสนอผลงานวิจัยของนิสิตบัณฑิตศึกษา ทำให้ได้รับประสบการณ์ในการนำเสนอสารนิพนธ์รวมถึงได้แลกเปลี่ยนความรู้ทางด้านเทคโนโลยีต่าง ๆ

นรเศรษฐ์ กรุดสมัย

สารบัญ

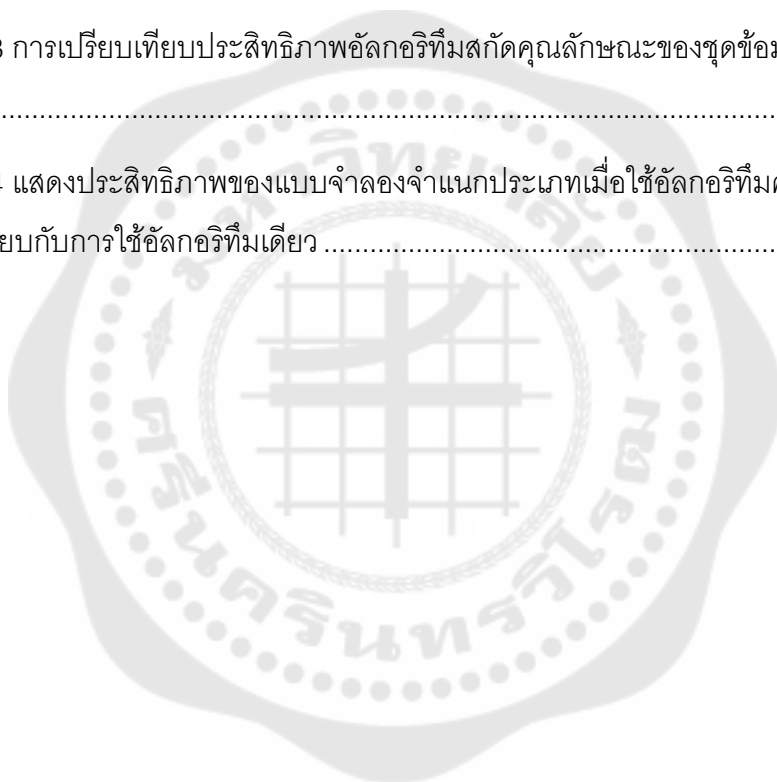
	หน้า
บทคัดย่อภาษาไทย	ง
บทคัดย่อภาษาอังกฤษ	จ
กิตติกรรมประกาศ.....	ฉ
สารบัญ	ช
สารบัญตาราง.....	ณ
สารบัญรูปภาพ	ญ
บทที่ 1 บทนำ.....	1
ภูมิหลัง	1
ความมุ่งหมายของการวิจัย.....	2
ความสำคัญของการวิจัย	3
ขอบเขตของการวิจัย	3
สมมติฐานในการวิจัย.....	3
บทที่ 2 ทบทวนวรรณกรรม.....	5
ประเภทของ Feature engineering (Zheng & Casari, 2018).....	6
ทฤษฎีอัลกอริทึมสกัดคุณลักษณะ Principal Component Analysis (PCA).....	6
ทฤษฎีอัลกอริทึมสกัดคุณลักษณะ Kernel PCA	7
ทฤษฎีอัลกอริทึมสกัดคุณลักษณะ Auto Encoder	8
ทฤษฎีอัลกอริทึมสกัดคุณลักษณะ t-Distributed Stochastic Neighbor Embedding (tSNE)	9
การจำแนกประเภทด้วยการใช้เทคนิคการเรียนรู้ของเครื่อง (Classification).....	11
ทฤษฎีอัลกอริทึม Logistic Regression	11

ทฤษฎีอัลกอริทึม Support Vector Machine (SVM).....	12
ทฤษฎีอัลกอริทึม Decision Tree.....	14
ทฤษฎีอัลกอริทึม Random Forest	15
การวัดประสิทธิภาพของอัลกอริทึม	16
งานวิจัยที่เกี่ยวข้อง	17
บทที่ 3 วิธีการดำเนินการวิจัย	19
ชุดข้อมูล	20
กำหนดอัลกอริทึมที่ใช้ทำการสกัดคุณลักษณะ	22
กำหนดอัลกอริทึมที่ใช้จำแนกข้อมูล	22
วิธีการดำเนินงานวิจัย	22
บทที่ 4 ผลการศึกษา	30
ผลลัพธ์การเปรียบเทียบประสิทธิภาพของแบบจำลองของชุดข้อมูล Credit Card Fraud Detection.....	30
ผลลัพธ์การเปรียบเทียบประสิทธิภาพของแบบจำลองของชุดข้อมูล Breast Cancer	33
ผลลัพธ์การเปรียบเทียบประสิทธิภาพของแบบจำลองของชุดข้อมูล Heart Disease	35
บทที่ 5 สรุป อภิปรายผล และข้อเสนอแนะ	37
สรุปผลการวิจัย.....	37
อภิปรายผล	40
ข้อเสนอแนะ	42
บรรณานุกรม	43
ประวัติผู้เขียน.....	45

สารบัญตาราง

หน้า

ตารางที่ 1 การเปรียบเทียบประสิทธิภาพอัลกอริทึมสกัดคุณลักษณะของชุดข้อมูล Credit Card Fraud Detection	38
ตารางที่ 2 การเปรียบเทียบประสิทธิภาพอัลกอริทึมสกัดคุณลักษณะของชุดข้อมูล Breast Cancer Wisconsin	39
ตารางที่ 3 การเปรียบเทียบประสิทธิภาพอัลกอริทึมสกัดคุณลักษณะของชุดข้อมูล Heart Disease	40
ตารางที่ 4 แสดงประสิทธิภาพของแบบจำลองจำแนกประเภทเมื่อใช้อัลกอริทึมคุณลักษณะแบบสองชั้นเทียบกับการใช้อัลกอริทึมเดียว	40



สารบัญรูปภาพ

หน้า

ภาพประกอบที่ 1 การหาองค์ประกอบหลักของ PCA	7
ภาพประกอบที่ 2 การทำงานของ Kernel PCA.....	8
ภาพประกอบที่ 3 โครงสร้างโครงข่ายประสาทเทียมของ Auto Encoder.....	9
ภาพประกอบที่ 4 การแมพिंगจุดข้อมูลจากมิติสูงสู่มิติต่ำของ tSNE.....	11
ภาพประกอบที่ 5 แสดงการจำแนกข้อมูลด้วยอัลกอริทึม SVM	13
ภาพประกอบที่ 6 แสดงการทำงานของ Kernel เพื่อแปลง Input Space เป็น Higher Dimensional Space	14
ภาพประกอบที่ 7 แสดงโครงสร้างของต้นไม้การตัดสินใจ	15
ภาพประกอบที่ 8 แสดงการทำงานของอัลกอริทึม Random Forest.....	16
ภาพประกอบที่ 9 ขั้นตอนการทดลอง.....	19
ภาพประกอบที่ 10 แสดงตัวอย่างชุดข้อมูล Credit Card Fraud Detection	20
ภาพประกอบที่ 11 แสดงตัวอย่างชุดข้อมูล Breast Cancer Wisconsin Data	21
ภาพประกอบที่ 12 แสดงตัวอย่างชุดข้อมูล Heart Disease UCI.....	21
ภาพประกอบที่ 13 แสดงขั้นตอนการทดลองโดยไม่ใช้อัลกอริทึมสกัดคุณลักษณะ	22
ภาพประกอบที่ 14 แสดงขั้นตอนการทดลองโดยใช้อัลกอริทึมสกัดคุณลักษณะ PCA.....	23
ภาพประกอบที่ 15 แสดงขั้นตอนการทดลองโดยใช้อัลกอริทึมสกัดคุณลักษณะ KPCA.....	24
ภาพประกอบที่ 16 แสดงขั้นตอนการทดลองโดยใช้อัลกอริทึมสกัดคุณลักษณะ AUTOENCODER	25
ภาพประกอบที่ 17 แสดงขั้นตอนการทดลองโดยใช้อัลกอริทึมสกัดคุณลักษณะ tSNE	26
ภาพประกอบที่ 18 แสดงขั้นตอนการทดลองโดยใช้อัลกอริทึมสกัดคุณลักษณะ PCA ร่วมกับ tSNE	27

ภาพประกอบที่ 19 แสดงขั้นตอนการทดลองโดยใช้อัลกอริทึมสกัดคุณลักษณะ KPCA ร่วมกับ tSNE	28
ภาพประกอบที่ 20 แสดงขั้นตอนการทดลองโดยใช้อัลกอริทึมสกัดคุณลักษณะ AUTOENCODER ร่วมกับ tSNE	29
ภาพประกอบที่ 21 แสดงผลการทำนายของชุดข้อมูล Credit Card Fraud Detection โดยใช้ อัลกอริทึมสกัดคุณลักษณะเดียว	31
ภาพประกอบที่ 22 แสดงผลการทำนายของชุดข้อมูล Credit Card Fraud Detection โดยใช้ อัลกอริทึมสกัดคุณลักษณะอื่นร่วมกับ tSNE	32
ภาพประกอบที่ 23 แสดงผลการทำนายของชุดข้อมูล Breast Cancer โดยใช้อัลกอริทึมสกัด คุณลักษณะเดียว	33
ภาพประกอบที่ 24 แสดงผลการทำนายของชุดข้อมูล Breast Cancer โดยใช้อัลกอริทึมสกัด คุณลักษณะอื่นร่วมกับ tSNE	34
ภาพประกอบที่ 25 แสดงผลการทำนายของชุดข้อมูล Heart Disease โดยใช้อัลกอริทึมสกัด คุณลักษณะเดียว	35
ภาพประกอบที่ 26 แสดงผลการทำนายของชุดข้อมูล Heart Disease โดยใช้อัลกอริทึมสกัด คุณลักษณะอื่นร่วมกับ tSNE	36

บทที่ 1

บทนำ

ภูมิหลัง

ในปัจจุบันเทคนิคการเรียนรู้ของเครื่องหรือ Machine Learning มีบทบาทกับชีวิตประจำวันของทุกคนอย่างหลีกเลี่ยงไม่ได้ไม่ว่าจะรู้ตัวหรือไม่รู้ตัวก็ตาม ซึ่งเทคนิคดังกล่าวเกิดจากความพร้อมหรือความเหมาะสมทางด้านวิทยาการคอมพิวเตอร์ คือมีพลังในการประมวลผลมากพอที่จะประมวลผลข้อมูลขนาดใหญ่ ทำให้มีการประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่องหรือ Machine Learning ในโปรแกรมประยุกต์ ในอุปกรณ์คอมพิวเตอร์รวมถึงสิ่งอำนวยความสะดวกทุกชนิด แต่ในความเป็นจริงแล้วการพัฒนาเทคนิคการเรียนรู้ของเครื่องมีความจำเป็นที่ต้องใช้ข้อมูลที่มากพอทั้งในแง่ปริมาณของคุณลักษณะข้อมูลซึ่งบ่งบอกลักษณะของข้อมูลและปริมาณข้อมูลทั้งหมดที่ใช้ประมวลผลเพื่อให้อัลกอริทึมสามารถที่จะค้นหารูปแบบหรือคำนวณผลลัพธ์ทางสถิติออกมาเพื่อในไปใช้งานต่อไป ดังนั้นนอกจากขั้นตอนในการเก็บรวบรวมข้อมูลแล้วขั้นตอนหนึ่งที่สำคัญในการพัฒนาเทคนิคการเรียนรู้ของเครื่องคือขั้นตอนการเตรียมข้อมูลก่อนการประมวลผลหรือ Data Pre-Processing ประกอบไปด้วยการทำข้อมูลให้เป็นมาตรฐาน (Standardizing Data) และการทำวิศวกรรมคุณลักษณะ (Feature Engineering) เนื่องจากข้อมูลที่ใช้ประมวลผลนั้นมีจำนวนตัวแปรหรือมิติข้อมูลจำนวนมาก สิ่งนี้เรียกว่า The Curse of Dimensionality ส่งผลต่อประสิทธิภาพของการประมวลผลเช่นใช้เวลานาน เกิดความเอนเอียงของผลลัพธ์ทางสถิติหรือแม้แต่การเข้าใจข้อมูลทำได้ยากจึงมีการใช้เทคนิคลดมิติของข้อมูลเพื่อแก้ปัญหาดังกล่าว

ในปัจจุบันมีการใช้เทคนิคการลดมิติข้อมูลอย่างแพร่หลาย การลดมิติของข้อมูลนั้นแบ่งออกเป็น 2 ประเภทได้แก่

1. การคัดเลือกคุณลักษณะ (Feature Selection) เป็นวิธีการเลือกคุณลักษณะข้อมูลมีอยู่ในชุดข้อมูล โดยจะเลือกเอาคุณลักษณะที่มีความสำคัญ ภายใต้ नियามบางอย่างออกมา และตัดคุณลักษณะที่มีความสำคัญน้อยออกไป ทำให้ลดจำนวนคุณลักษณะลงได้

2. การสกัดคุณลักษณะ (Feature Extraction) เป็นวิธีการสกัดคุณลักษณะข้อมูลโดยการสร้างคุณลักษณะขึ้นมาใหม่ จากการทำผลรวมแบบเชิงเส้นของคุณลักษณะข้อมูลที่มีอยู่เดิม แล้วคัดเลือกเอา คุณลักษณะที่มีความสำคัญที่สุดออกมา ซึ่งในวิธีการนี้ รูปแบบของข้อมูลจะเปลี่ยนแปลงไปจากเดิม ซึ่งแตกต่างจากการคัดเลือกคุณลักษณะ ที่คุณลักษณะจะไม่เปลี่ยนแปลงไปจากเดิม เพียงแต่ตัดเลือกบางค่ามาเท่านั้น

ในการศึกษานี้ เราจะมุ่งความสนใจไปที่การสกัดคุณลักษณะ โดยอัลกอริทึมสกัดคุณลักษณะที่เป็นที่นิยมในหลายปีที่ผ่านมาคืออัลกอริทึม tSNE (van der Maaten, & Hinton, 2008) ได้ถูกนำไปใช้กับงานทางวิทยาศาสตร์หลากหลายด้านเพื่อลดจำนวนมิติข้อมูล ไม่ว่าจะเป็นทางด้าน Bio Informatics (Kobak & Berens, 2019) ด้านการวิเคราะห์ทางธุรกิจ เป็นต้น โดยอัลกอริทึม tSNE นี้มีความสามารถในการลดมิติข้อมูลจากมิติสูงลงสู่มิติต่ำ เพื่อใช้ในการสร้างมโนทัศน์ (visualization) ของข้อมูล โดยการลดมิติให้เหลือ 2 – 3 มิติ

ทว่าการใช้อัลกอริทึม tSNE กับข้อมูลเพื่อลดจำนวนมิติลงนั้น หากลดมิติของข้อมูลลงจากข้อมูลที่มีมิติจำนวนมาก ๆ ไปเป็นข้อมูลที่มีขนาด 2-3 มิติในคราวเดียว จะได้ผลไม่ดีนัก โดยจะส่งผลให้ข้อมูลที่ถูกลดมิติลง มีประสิทธิภาพการจำแนกข้อมูลลดลงอย่างมาก เนื่องจากอัลกอริทึม tSNE ไม่สามารถรักษาคุณสมบัติโครงสร้างโดยรวมของข้อมูล (Global Structure) ไว้ได้ ทำให้ข้อมูลที่ได้อัลกอริทึมมีประสิทธิภาพในการจำแนกข้อมูลไม่ดี เนื่องจากข้อมูลที่เคยอยู่ห่างกัน ขณะอยู่ในมิติสูงอาจไม่ได้อยู่ห่างกันเช่นเดิม เมื่อเราลดมิติลงให้เหลือมิติเพียง 2 – 3 มิติ

ซึ่งวิธีหนึ่งในการแก้ปัญหานี้คือ การใช้วิธีลดมิติแบบสองขั้น กล่าวคือจะมีการใช้อัลกอริทึมลดมิติข้อมูลขั้นหนึ่งก่อนเพื่อลดจำนวนมิติลงไปขั้นหนึ่ง จากนั้นจึงนำข้อมูลที่ได้อัลกอริทึมลดมิติขั้นที่สอง การทำเช่นนี้จะช่วยลดปัญหาของการสูญเสียโครงสร้างโดยรวมของข้อมูลไว้ได้

ทั้งนี้ได้มีงานวิจัยบางส่วน ที่ศึกษาการใช้งานอัลกอริทึมลดมิติอื่นร่วมกับ อัลกอริทึม tSNE เช่น (Kobak & Berens, 2019) ซึ่งใช้อัลกอริทึม PCA ร่วมกับอัลกอริทึม tSNE ทว่านอกจากอัลกอริทึม PCA แล้วยังมีอัลกอริทึมลดมิติอื่น ๆ อีกมากมาย จึงทำให้เป็นที่มาของการศึกษาในครั้งนี้มีจุดประสงค์เพื่อศึกษาว่าใช้อัลกอริทึมลดมิติแบบอื่นร่วมกับ tSNE นั้นจะส่งผลเหมือนหรือแตกต่างจากการใช้อัลกอริทึม PCA อย่างไร นอกจากนี้ เรายังสนใจศึกษาว่าควรใช้อัลกอริทึมลดมิติชนิดใดร่วมกับอัลกอริทึม tSNE จึงจะมีประสิทธิภาพดีที่สุด

ความมุ่งหมายของการวิจัย

ในการวิจัยครั้งนี้ผู้วิจัยได้ตั้งความมุ่งหมายไว้ดังนี้

1. ศึกษาแนวคิดและวิธีการทำงานของอัลกอริทึมที่ใช้สกัดคุณลักษณะโดยเฉพาะอัลกอริทึม tSNE
2. ศึกษาการลดมิติของข้อมูลให้สามารถแสดงได้ในรูปแบบ 2 มิติ

3. ศึกษาและเปรียบเทียบประสิทธิภาพอัลกอริทึมสกัดคุณลักษณะที่นำมาใช้งานร่วมกันเป็นขั้นที่หนึ่งเมื่อทำงานร่วมกับอัลกอริทึม tSNE โดยวัดจากประสิทธิภาพในการจำแนกข้อมูลด้วยอัลกอริทึมจำแนกประเภท (Classification)

ความสำคัญของการวิจัย

ผลที่ได้จากงานวิจัยนี้สามารถนำไปใช้เป็นแนวทางในการเลือกใช้อัลกอริทึมสกัดคุณลักษณะ นำไปทำงานร่วมกับอัลกอริทึม tSNE ในการลดมิติของข้อมูล โดยไม่ทำให้สูญเสียประสิทธิภาพในการจำแนกประเภทไป โดยอัลกอริทึมสกัดคุณลักษณะที่ได้รับความนิยมถูกเลือกใช้ในครั้งนี้อัลกอริทึม PCA, Kernel PCA, AUTOENCODER และ tSNE ในการศึกษาได้นำแบบจำลองที่ใช้จำแนกประเภทที่ได้รับความนิยมได้แก่ Logistic Regression, Support Vector Machine, Decision Tree และ Random Forest มาใช้ในการวัดประสิทธิภาพ โดยใช้คะแนนความแม่นยำของผลลัพธ์ในการจำแนกประเภท (Accuracy Score) เป็นตัวแสดงสะท้อนประสิทธิภาพของอัลกอริทึมสกัดคุณลักษณะที่ถูกนำมาศึกษา

ขอบเขตของการวิจัย

งานวิจัยนี้มุ่งเน้นที่จะศึกษาผลของการใช้อัลกอริทึมสกัดคุณลักษณะร่วมกันทำงานกับอัลกอริทึม tSNE โดยมีขอบเขตของการวิจัยดังนี้

1. ใช้ชุดข้อมูลจำนวน 3 ชุด ประกอบด้วย Credit Card Fraud Detection, Breast Cancer Wisconsin Data และ Heart Disease UCI
2. ใช้อัลกอริทึมสกัดคุณลักษณะได้แก่ PCA, Kernel PCA, AUTOENCODER และ tSNE
3. ใช้อัลกอริทึมจำแนกประเภทได้แก่ Logistic Regression, Support Vector Machine, Decision Tree และ Random Forest
4. การแสดงผลข้อมูลโดยใช้อัลกอริทึมสกัดคุณลักษณะให้อยู่ใน 2 มิติ
5. เปรียบเทียบประสิทธิภาพการจำแนกข้อมูลโดยใช้ Accuracy Score

สมมติฐานในการวิจัย

1. การใช้อัลกอริทึมสกัดคุณลักษณะอื่น ร่วมทำงานกับอัลกอริทึม tSNE แบบสองชั้นจะช่วยเพิ่มประสิทธิภาพการทำงาน ในแง่การลดมิติข้อมูลโดยยังคงรักษาโครงสร้างโดยรวมของข้อมูลเพื่อนำข้อมูลไปใช้ในการจำแนกประเภทได้อย่างมีประสิทธิภาพ

2. ในการใช้อัลกอริทึมสกัดคุณลักษณะที่ต่างกัันมาร่วมทำงานกับอัลกอริทึม tSNE จะให้ประสิทธิภาพที่ต่างกััน ดังนั้นจึงต้องการค้นหาอัลกอริทึมสกัดคุณลักษณะที่มีความเหมาะสม ที่สามารถทำงานร่วมกับอัลกอริทึม tSNE ได้อย่างมีประสิทธิภาพมากที่สุด



บทที่ 2

ทบทวนวรรณกรรม

ในการทำ machine learning มีอัลกอริทึมเป็นจำนวนมากที่ทำงานได้ดีและรวดเร็วกับข้อมูลที่มีจำนวนมิติไม่มากนัก แต่เมื่อข้อมูลมีจำนวนมิติหรือคุณลักษณะเป็นจำนวนมาก จะเกิดปรากฏการณ์ที่เรียกว่า Curse of Dimensionality (Altman & Krzywinski, 2018) ซึ่งทำให้การประมวลผลข้อมูลใช้เวลาและพื้นที่หน่วยความจำอย่างมาก

นอกจากนี้ข้อมูลที่มีมิติมากเกินไปในบางมิติอาจมีข้อมูลที่ซ้ำซ้อนหรือมีความสัมพันธ์กับข้อมูลในมิติอื่น หรือแม้แต่ว่ามีความสัมพันธ์แบบเชิงเส้นกับในบางมิติได้ (co-Linearity) ทั้งนี้ในบางมิติอาจมีการเก็บข้อมูลที่ผิดพลาด ทำให้เกิดสัญญาณรบกวนในข้อมูล (Noise) ทำให้ไม่สามารถประมวลผลข้อมูลได้อย่างมีประสิทธิภาพ ดังนั้นในข้อมูลที่มีมิติเป็นจำนวนมากจึงนิยมลดขนาดของมิติลงก่อนที่จะนำไปใช้วิเคราะห์ในขั้นตอนต่อไป โดยสามารถสรุปประโยชน์ในการลดมิติของข้อมูลได้ดังนี้

1. เมื่อมิติของข้อมูลลดลงจะใช้พื้นที่ในการเก็บข้อมูลลดลงส่งผลให้ใช้เวลาในการประมวลผลหรือฝึกฝน model น้อยลง
2. ในบางอัลกอริทึมการที่มีมิติข้อมูลจำนวนมากทำให้ประสิทธิภาพลดลง การลดมิติข้อมูลช่วยให้อัลกอริทึมมีประสิทธิภาพดีขึ้น
3. เพื่อลดมิติของข้อมูลให้สามารถอธิบายได้ในรูปแบบ 2 – 3 มิติ
4. เพื่อลดความสัมพันธ์ระหว่างคุณลักษณะโดยจะยังคงประสิทธิภาพของแบบจำลองดีเช่นเดิมหรือดีขึ้น
5. ช่วยลดสัญญาณรบกวนในคุณลักษณะบางส่วนได้

การลดมิติสามารถแบ่งออกได้เป็นสองรูปแบบคือ การเลือกคุณลักษณะ (Feature Selection) และ การสกัดคุณลักษณะ (Feature Extraction) ซึ่งมีรายละเอียดดังนี้ โดยอัลกอริทึมในการเลือกคุณลักษณะจะทำหน้าที่คัดกรอง เลือกเอาคุณลักษณะที่สำคัญและมีข้อมูลอยู่มากเก็บไว้ และคัดแยกคุณลักษณะที่มีข้อมูลอยู่น้อยออกไปจากข้อมูล ส่วนการสกัดคุณลักษณะจะเป็นการสร้างคุณลักษณะใหม่ จากคุณลักษณะที่มีอยู่เดิมโดยใช้อัลกอริทึมต่าง ๆ ในการคำนวณหาความสำคัญของคุณลักษณะเดิม

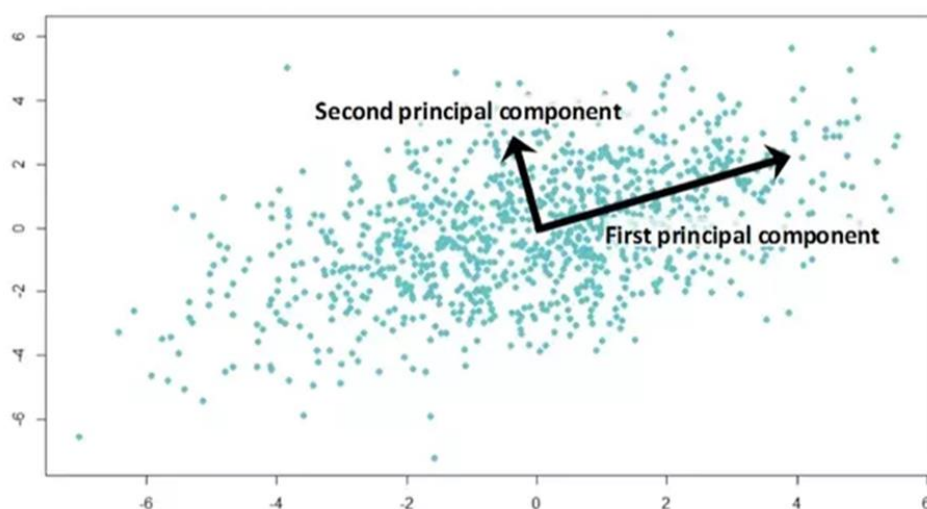
ประเภทของ Feature engineering (Zheng & Casari, 2018)

1. การเลือกคุณลักษณะ (Feature selection) เป็นการคัดกรองคุณลักษณะที่ไม่เกี่ยวข้องหรือซ้ำซ้อนในชุดข้อมูล โดยการทำ feature selection จะเลือกคุณลักษณะที่มีลักษณะเด่นในชุดข้อมูลและทิ้งคุณลักษณะที่ไม่สำคัญไป โดยจะไม่มีเปลี่ยนแปลงรูปแบบของแต่ละคุณลักษณะเลย (Novakovic, 2016)

2. การแยกคุณลักษณะ (Feature extraction) เป็นการสร้างคุณลักษณะใหม่โดยมีจำนวนน้อยกว่า (Heaton, 2016) คุณลักษณะเดิมซึ่งยังคงคุณลักษณะของข้อมูลที่เป็นประโยชน์ โดยการสร้างคุณลักษณะใหม่นี้ จะมีวิธีการที่แตกต่างกัน เช่นสร้างจากการรวมเชิงเส้น (Linear Combination) หรือผ่านการแมปไปยังแบบไม่เชิงเส้นชนิดอื่น ๆ ได้ ในการศึกษานี้จะทำการศึกษากับอัลกอริทึมสำหรับแยกคุณลักษณะ โดยในหัวข้อถัดไปได้้นำเสนออัลกอริทึมที่นิยมใช้ในการแยกคุณลักษณะดังต่อไปนี้

ทฤษฎีอัลกอริทึมสกัดคุณลักษณะ Principal Component Analysis (PCA)

PCA เป็นเทคนิคประเภท unsupervised ที่ช่วยแยกคุณลักษณะจากคุณลักษณะที่มีอยู่เดิมโดยจะทำการ project ลงในชุดข้อมูล ค้นหาความแปรปรวนของคุณลักษณะที่มากที่สุดซึ่งคุณลักษณะใหม่นี้เรียกว่า Principal Components โดยจะเรียงลำดับคุณลักษณะที่สามารถอธิบายความแปรปรวนของคุณลักษณะ (Explained Variance) ได้มากที่สุด เรียกว่า First Principal Component และลำดับ ถัดไป เรียกว่า Second Principal Component

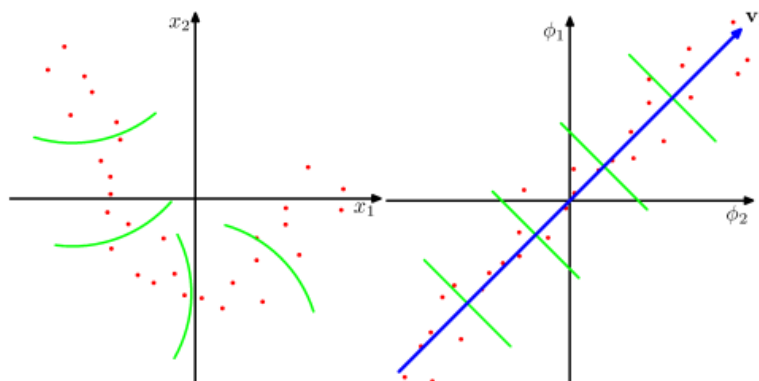


ภาพประกอบที่ 1 การหาองค์ประกอบหลักของ PCA

ที่มา ๑: Analyticsvidhya (2016). PCA: A Practical Guide to Principal Component Analysis in R & Python. Retrieved from <https://www.analyticsvidhya.com/blog/2016/03/pca-practical-guide-principal-component-analysis-python/>

ทฤษฎีอัลกอริทึมสกัดคุณลักษณะ Kernel PCA

ทำงานเช่นเดียวกับ PCA แต่มีการใช้ Kernel Trick เพื่อเพิ่มมิติของข้อมูลแล้วจึงค้นหาองค์ประกอบหลักจากข้อมูล



ภาพประกอบที่ 2 การทำงานของ Kernel PCA

ที่มา : Bishop, C. M. (2006). Pattern Recognition and Machine Learning (Information Science and Statistics): Springer-Verlag.

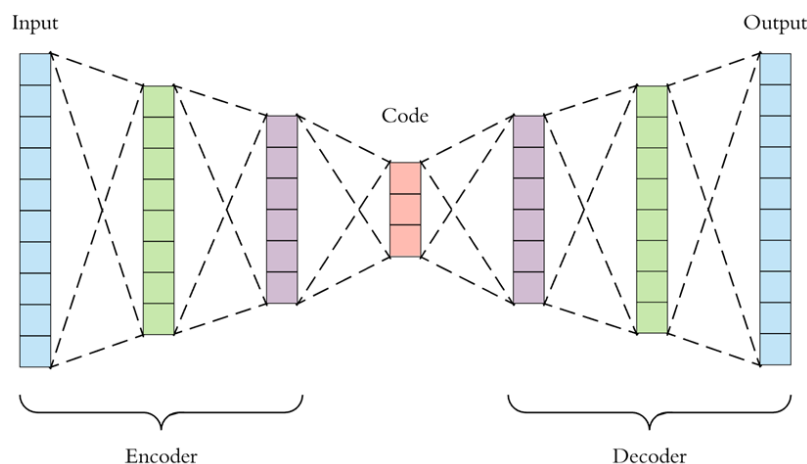
ทฤษฎีอัลกอริทึมสกัดคุณลักษณะ Auto Encoder

Auto Encoder เป็นโครงสร้างของโครงข่ายประสาทเทียมที่สามารถที่จะเรียนรู้การบีบอัดข้อมูลที่มีประสิทธิภาพ เริ่มต้นด้วยเข้ารหัสข้อมูลต่อจากนั้นจะทำการหาวิธีที่จะสร้างข้อมูลกลับจากข้อมูลที่ถูกบีบอัดนั้นโดยพยายามสร้างให้ข้อมูลนั้นกลับมาเหมือนข้อมูล input มากที่สุด สรุปคือ Auto Encoder ถูกออกแบบให้ลดมิติข้อมูลโดยที่จะเรียนรู้วิธีกำจัด Noise ในข้อมูล

Auto Encoder ประกอบด้วย 4 ส่วนสำคัญคือ

1. Encoder: เป็นส่วนที่เรียนรู้ที่จะลดมิติข้อมูลจาก input และบีบอัดข้อมูล input เข้าสู่รูปแบบการเข้ารหัส
2. Latent หรือ Code: เป็นชั้นข้อมูลที่ถูกบีบอัดซึ่งถูกส่งต่อมาจากชั้น Encoder โดยชั้นข้อมูลนี้จะมีมิติของข้อมูลน้อยที่สุด
3. Decoder: ในขั้นนี้ Model จะทำเรียนรู้ที่จะสร้างข้อมูลใหม่เพื่อให้ใกล้เคียงกับข้อมูลที่ได้รับเข้ามาจากชั้น encoder ให้มากที่สุดเท่าที่เป็นไปได้
4. Reconstruction Loss: เป็นวิธีที่วัดว่าชั้น Decoder ทำงานได้ดีเพียงใดจากข้อมูล Output เทียบกับ Input

โดยการ training จะใช้วิธีการ Back Propagation เพื่อที่จะลดค่า Reconstruction Loss ให้น้อยที่สุด



ภาพประกอบที่ 3 โครงสร้างโครงข่ายประสาทเทียมของ Auto Encoder

ที่มา: Karpinski, M., Khoma, V. V., Dudkevych, V., Khoma, Y., & Sabodashko, D. (2018). Autoencoder Neural Networks for Outlier Correction in ECG- Based Biometric Identification. 2018 IEEE 4th International Symposium on Wireless Systems within the International Conferences on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS-SWS), 210-215.

ทฤษฎีอัลกอริทึมสกัดคุณลักษณะ t-Distributed Stochastic Neighbor Embedding (tSNE)

tSNE เป็นวิธีที่นิยมใช้ในการแสดงผลข้อมูลที่มีจำนวนมิติสูงอย่างแพร่หลาย จุดประสงค์ของ tSNE คือการใช้ชุดของจุดที่อยู่ในข้อมูลมิติสูงและหาตัวแทนที่แสดงถึงจุดเหล่านั้นในมิติที่ต่ำลง โดยอัลกอริทึมนี้เป็น Non – Linear ซึ่งแตกต่างจาก PCA ซึ่งเป็นการใช้เทคนิค Linear Combination โดยจะปรับข้อมูลตามความแตกต่างกันใน Region ต่าง ๆ ซึ่งความแตกต่างเหล่านั้นเป็นข้อมูลสำคัญในการนำไปใช้แสดงผล tSNE จะทำการคำนวณความคล้ายคลึงกันระหว่างคู่ของจุดข้อมูลในพื้นที่มิติสูงและพื้นที่ในมิติต่ำ จากนั้นจะทำการปรับความคล้ายคลึงในมิติ 2 ระดับนี้ด้วย Cost Function tSNE จะสร้างการกระจายตัวของความน่าจะเป็น (Probability Distribution) โดยใช้ Gaussian Distribution กำหนดความสัมพันธ์ระหว่างจุดข้อมูลในมิติสูง และ tSNE จะใช้ Student t-distribution เพื่อสร้างความน่าจะเป็นแบบกระจายตัวในมิติต่ำเพื่อป้องกันการกระจุกตัวหรือทับซ้อนของจุดข้อมูลในมิติต่ำซึ่งเป็นผลมาจาก Curse of Dimensionality

ข้อดี

1. ใช้ Gaussian kernel กำหนดขอบเขตระหว่างโครงสร้าง Local และ Global ของชุดข้อมูล

2. ในการแสดงผลข้อมูลมิติต่ำของจุดข้อมูลใกล้เคียงและไกลจะมีความสำคัญเท่ากัน

3. ขนาดของกลุ่มจุดข้อมูลขึ้นอยู่กับความหนาแน่นของข้อมูล

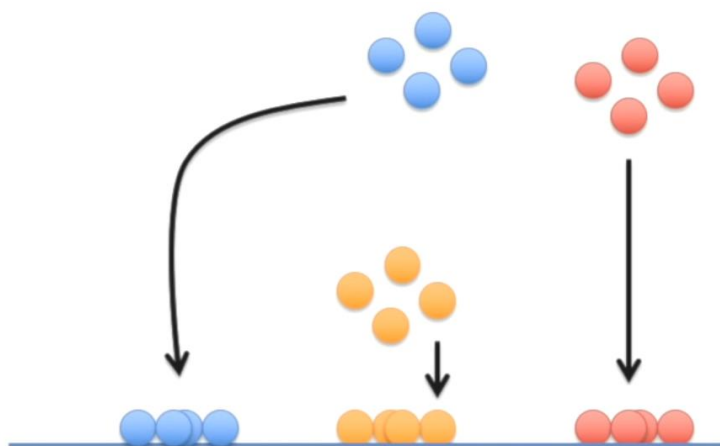
ข้อสังเกต

1. tSNE ยังมีปัญหาในการลดข้อมูลที่เหลือมากกว่า 3 มิติ

2. tSNE ลดมิติข้อมูลโดยคำนึงถึงคุณสมบัติในระดับ Local

3. tSNE จะเกิดปัญหา Curse of the Intrinsic Dimensionality of the Data คือ เมื่อมีจำนวนมิติที่ต้องการคำนวณจำนวนมากต้องใช้เวลาในการประมวลผลสูงมาก

4. tSNE เพิ่มประสิทธิภาพด้วยการใช้ Gradient Descent ใช้ Cost Function แบบ Non – Convex จึงมีโอกาที่ tSNE จะหยุดหาจุดต่ำสุดที่ Local ดังนั้นต้องใช้วิธีต่าง ๆ ที่จะหลีกเลี่ยงปัญหานี้ (Kurita, 2018)



ภาพประกอบที่ 4 การแมพปิ้งจุดข้อมูลจากมิติสูงสู่มิติต่ำของ tSNE

ที่มา : Starmer, J. (Producer). StatQuest: t-SNE, Clearly Explained.

<https://youtu.be/NEaUSP4YerM>

การจำแนกประเภทด้วยการใช้เทคนิคการเรียนรู้ของเครื่อง (Classification)

Classification เป็นเทคนิคหนึ่งของการเรียนรู้ของเครื่องแบบมีผู้สอน (Supervised) โดยกระบวนการเรียนรู้นั้นจำเป็นต้องมีข้อมูลคำตอบ ซึ่งจะแบ่งข้อมูลคำตอบออกเป็น Class ต่าง ๆ สำหรับหลักการทำงานของแบบจำลองจะแบ่งข้อมูลออกเป็น 2 ส่วน คือชุดข้อมูลการเรียนรู้และชุดข้อมูลทดสอบ ซึ่งในงานวิจัยนี้แบ่งข้อมูลดังกล่าวในอัตราส่วน 80 : 20 หรือตามความเหมาะสม จากนั้นนำข้อมูลชุดการเรียนรู้ป้อนเข้าสู่ระบบเพื่อเข้าสู่ขั้นตอนการเรียนรู้โดยใช้อัลกอริทึมที่กำหนด จนกระทั่งได้ผลลัพธ์ออกมาเป็นแบบจำลองการทำนาย จากนั้นนำข้อมูลชุดทดสอบซึ่งเป็นข้อมูลที่เตรียมไว้สำหรับทดสอบประสิทธิภาพการทำงานของแบบจำลองมาใช้ในการทำนายผลและพิจารณาผลลัพธ์ที่ได้ โดยในงานวิจัยนี้ได้ใช้เทคนิคในการจำแนกกลุ่มข้อมูล ดังนี้

ทฤษฎีอัลกอริทึม Logistic Regression

การวิเคราะห์การถดถอยโลจิสติก (Logistic Regression) เป็นอัลกอริทึมที่ใช้ในการวิเคราะห์ข้อมูลเชิงคุณภาพ (Qualitative data) ซึ่งลักษณะของข้อมูลนั้นเป็นประเภทหมวดหมู่ การวิเคราะห์การถดถอยโลจิสติกมักถูกใช้กับงานด้านการจำแนกประเภทข้อมูล เพื่อทำนายความน่าจะเป็นของการเกิดของเหตุการณ์ต่าง ๆ โดยสามารถแบ่งประเภทการวิเคราะห์ได้ 3 ประเภทหลักๆ ดังนี้

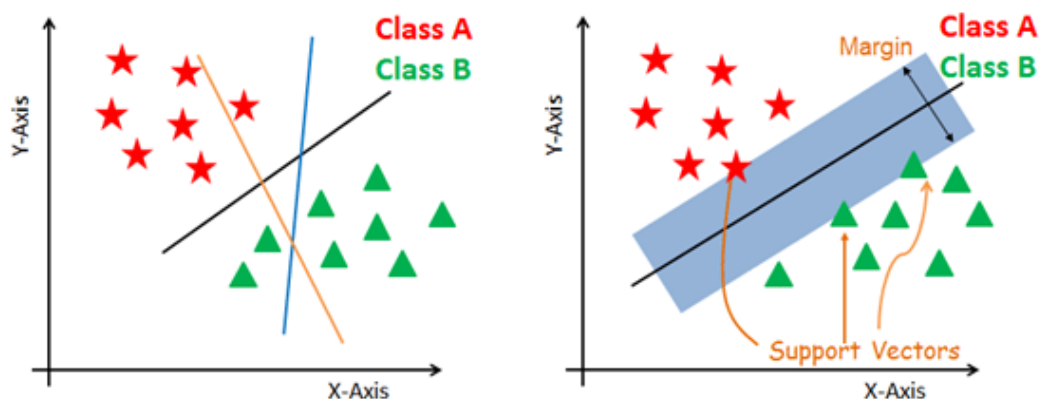
1. การวิเคราะห์การถดถอยโลจิสติกแบบไบนารี (Binary Logistic Regression Analysis) คือการวิเคราะห์ข้อมูลโดยที่ค่าของคำตอบ (Labels) มีความน่าจะเป็น 2 ประเภท

2. การวิเคราะห์การถดถอยโลจิสติกพหุกลุ่ม (Multinomial Logistic Regression Analysis) เป็นการวิเคราะห์ข้อมูลโดยที่ค่าของคำตอบมีความน่าจะเป็นมากกว่า 2 ประเภท ซึ่งเป็นค่าที่ไม่สามารถจัดลำดับได้

3. การวิเคราะห์การถดถอยโลจิสติกเชิงลำดับ (Ordinal Logistic Regression) เป็นการวิเคราะห์ข้อมูลโดยที่ค่าของคำตอบมีความน่าจะเป็นมากกว่า 2 ประเภท ซึ่งเป็นค่าที่สามารถจัดลำดับได้ โดยหลักการทำงานจะต้องมีการกำหนดขอบเขตการตัดสินใจ (Decision Boundaries) เพื่อระบุว่าข้อมูลที่ทำนายควรจะอยู่ประเภทใด จากนั้นจะทำการประมาณค่าสัมประสิทธิ์การถดถอยด้วยเทคนิคต่าง ๆ โดยเทคนิคที่นิยมได้แก่ Maximum Likelihood เป็นต้น (Chandrayan, 2015)

ทฤษฎีอัลกอริทึม Support Vector Machine (SVM)

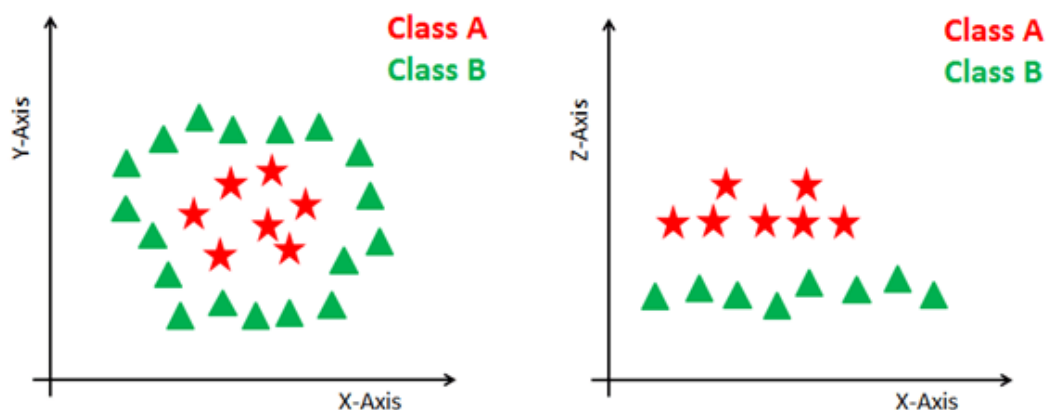
SVM เป็นอัลกอริทึมหนึ่งสำหรับงานด้านการจำแนกประเภทข้อมูลที่สามารถทำงานได้อย่างมีประสิทธิภาพ แม้ในกรณีที่ชุดข้อมูลมีจำนวนมิติ (Dimensions) มากกว่าจำนวนตัวอย่าง (Samples) โดยสามารถจัดการทั้งปัญหาเชิงเส้นตรง (Linear Problem) และปัญหาที่ไม่ใช่เชิงเส้นตรง (Non-Linear Problem) สำหรับหลักการทำงานจะเริ่มจากการวางข้อมูลบน Feature Space และสร้างเส้นแบ่งกลุ่มข้อมูล (Hyperplane) เพื่อแบ่งข้อมูลออกเป็นกลุ่ม ๆ โดย Hyperplane ที่ดีนั้นจะต้องมี Margin กว้างที่สุดหรือที่เรียกว่า Maximum Margin ที่สามารถจำแนกข้อมูลแต่ละประเภทออกจากกันได้อย่างชัดเจน ภาพประกอบที่ 5 แสดงการจำแนกข้อมูล 2 ประเภทโดยที่สมาชิกของข้อมูลแต่ละประเภทที่อยู่ใกล้ Hyperplane มากที่สุดเรียกว่า Support Vector มีระยะห่างจาก Hyperplane ที่กว้างเท่า ๆ กัน อาจเรียกได้ว่ามีขนาดของ Margin ที่กว้างเท่า ๆ กัน



ภาพประกอบที่ 5 แสดงการจำแนกข้อมูลด้วยอัลกอริทึม SVM

ที่มา: Navlani, A. (2019). Support Vector Machines with Scikit-learn. Retrieved from <https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python#svm>

อีกหนึ่งความสามารถของ SVM คือการจัดการกับปัญหากรณี Non-Linear Problem คือการจัดการกับข้อมูลที่ไม่สามารถใช้ Linear Hyperplane ได้ ดังแสดงในภาพประกอบที่ ซึ่งอธิบายหลักการทำงานของ SVM ภาพด้านซ้ายแสดงภาพของข้อมูลจำนวน 2 กลุ่มที่ไม่สามารถทำการจำแนกได้ด้วยเส้นตรง ภาพทางด้านขวาแสดงการใช้เทคนิค Kernel มาช่วยเพิ่มความสามารถในการจำแนกข้อมูล โดยการแปลงพื้นที่ข้อมูลที่นำเข้า (Input Space) เป็นพื้นที่ของมิติที่สูงกว่า ซึ่งแต่ละ Data Points จะถูก Plot ลงบนแกน X และแกน Z (โดยที่ Z คือ ผลรวมยกกำลังสองของ X และ Y : $Z = X^2 + Y^2$) (Navlani, 2019)

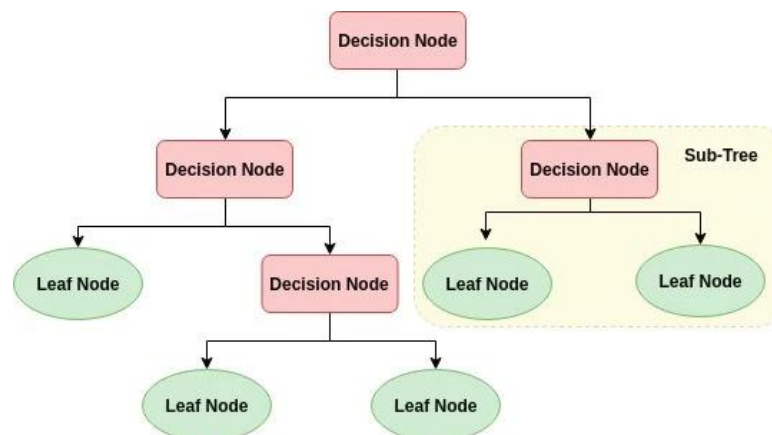


ภาพประกอบที่ 6 แสดงการทำงานของ Kernel เพื่อแปลง Input Space เป็น Higher Dimensional Space

ที่มา: Navlani, A. (2019). Support Vector Machines with Scikit-learn. Retrieved from <https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python#svm>

ทฤษฎีอัลกอริทึม Decision Tree

เป็นอัลกอริทึมที่ใช้ในการแยกประเภทด้วยวิธีการใช้ต้นไม้การตัดสินใจ โดยแต่ละโหนดจะเป็นตัวแทนของคุณลักษณะของข้อมูล กิ่งแทนกฎที่ใช้ตัดสินใจส่วนใบแทนผลลัพธ์ของประเภทที่ต้องการจำแนกโดยส่วนที่อยู่บนสุดเรียกว่าราก อัลกอริทึมจะเรียนรู้ผ่านส่วนของคุณลักษณะ ภาพประกอบที่ แสดงให้เห็นถึงโครงสร้างของต้นไม้การตัดสินใจ

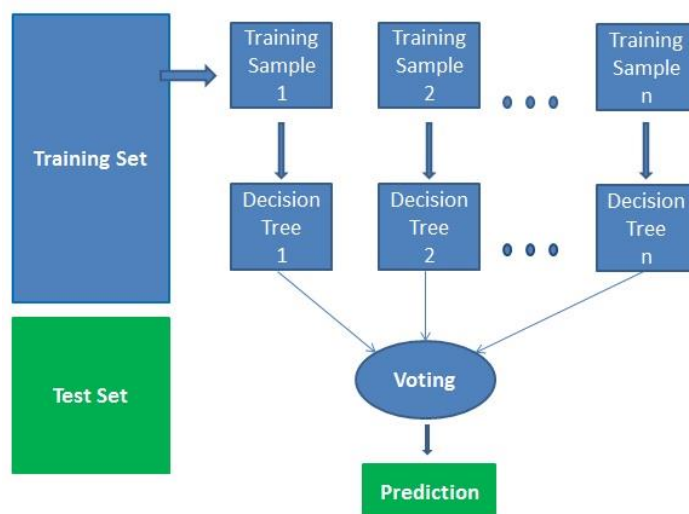


ภาพประกอบที่ 7 แสดงโครงสร้างของต้นไม้การตัดสินใจ

ที่มา: Navlani, A. (2018). Decision Tree Classification in Python. Retrieved from <https://www.datacamp.com/community/tutorials/decision-tree-classification-python>

ทฤษฎีอัลกอริทึม Random Forest

Random Forest เป็นเทคนิคที่พัฒนามาจากเทคนิค Decision Tree สำหรับงานด้านการวิเคราะห์เชิงถดถอยและการจำแนกกลุ่มข้อมูล สำหรับ Random Forest Classification นั้น จะสร้างแบบจำลองการทำนายในรูปแบบของต้นไม้การตัดสินใจหลาย ๆ ต้น (Ensemble of Decision Trees) เพื่อช่วยในการทำนายผลลัพธ์ โดยค่า Correlation ระหว่างต้นไม้การตัดสินใจแต่ละต้นจะถูกสร้างให้มีความเป็นอิสระต่อกัน (Independent) โดยจะสุ่มเลือกชุดข้อมูลการเรียนรู้แบบ Bootstrap จากชุดข้อมูลการเรียนรู้เดิม ซึ่งชุดข้อมูลที่สุ่มเลือกนั้นมีโอกาสที่จะถูกสุ่มเลือกซ้ำ โดย ต้นไม้การตัดสินใจแต่ละต้นจะสุ่มเลือกชุดข้อมูลคุณลักษณะที่ใช้ในการทำนายทำให้ต้นไม้การตัดสินใจแต่ละต้นมีลักษณะไม่เหมือนกัน จากนั้นจะใช้วิธีการ Majority Vote ช่วยในการตัดสินใจผลการทำนาย โดยนำผลการทำนายของต้นไม้การตัดสินใจแต่ละต้นมารวมกันตัดสินใจ จากนั้นจึงเลือกผลลัพธ์การทำนายที่ได้รับผลโหวตมากที่สุด



ภาพประกอบที่ 8 แสดงการทำงานของอัลกอริทึม Random Forest

ที่มา : Avinash Navlani. (2018). Understanding Random Forest s Classifiers in Python. Retrieved from <https://www.datacamp.com/community/tutorials/random-forestsclassifier-python>

การวัดประสิทธิภาพของอัลกอริทึม

ในส่วนของการทำนายแบบการจำแนกประเภท (Classification) ในงานวิจัยนี้จะใช้ค่า Accuracy ในการวัดประสิทธิภาพของอัลกอริทึม

Accuracy คือการวัดค่าความถูกต้องโดยรวมของระบบระหว่างค่าจริงและค่าการทำนาย หากค่า Accuracy มีค่ามากหมายถึงค่าการทำนายนั้นสามารถทำนายได้ถูกต้องใกล้เคียงกับค่าจริง ดังสมการ

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

จากสมการ

Accuracy คือ ค่าความถูกต้อง

TP	คือ ค่าการทำนายที่ทำนายว่าจริงซึ่งตรงกับค่าจริง
TN	คือ ค่าการทำนายที่ทำนายว่าไม่จริงซึ่งตรงกับค่าจริง
FP	คือ ค่าการทำนายที่ทำนายว่าไม่จริงซึ่งไม่ตรงกับค่าจริง
FN	คือ ค่าการทำนายที่ทำนายว่าจริงซึ่งไม่ตรงกับค่าจริง

งานวิจัยที่เกี่ยวข้อง

ทบทวนวรรณกรรมและงานวิจัยที่เกี่ยวข้องเกี่ยวกับทฤษฎีที่ใช้ในงานวิจัยนี้

1. บทความวิจัยเรื่อง Visualizing Data using tSNE (van der Maaten & Hinton, 2008) งานวิจัยนี้ได้ทำการเสนอเทคนิคการแสดงผลข้อมูลมิติสูงแบบใหม่เรียกว่า tSNE โดยให้อยู่ในรูปแบบ 2 – 3 มิติ งานวิจัยนี้มีพื้นฐานจากเทคนิค Stochastic Neighbor Embedding (Hinton and Roweis, 2002) ซึ่งทำการเพิ่มประสิทธิภาพเพื่อให้แสดงผลได้ดีขึ้นโดยลดการกระจุกตัวของจุดข้อมูลเมื่อแสดงผล ซึ่งสาเหตุมาจากเมื่อต้องการแสดงผลข้อมูลมิติสูงยากที่จะแยกจุดข้อมูลที่ใกล้เคียงออกจากจุดข้อมูลที่ห่าง งานวิจัยนี้อ้างว่า tSNE เป็นเทคนิคที่ดีกว่าเทคนิคอื่น ๆ ในการแสดงผลข้อมูลระดับต่าง ๆ ให้อยู่ในแผนที่เดียวกัน งานวิจัยนี้ได้แสดงตัวอย่างการแสดงผลชุดข้อมูลขนาดใหญ่โดยเปรียบเทียบกับหลาย ๆ เทคนิคคือ Sammon mapping, Isomap, และ Locally Linear Embedding ผู้วิจัยได้ศึกษาการทำงานของอัลกอริทึม tSNE เพื่อหาข้อดีและข้อจำกัดและพบว่าอัลกอริทึม tSNE ยังไม่มีประสิทธิภาพที่ดีนักเนื่องจากการลดมิติทำให้สูญเสียลักษณะที่สำคัญของข้อมูลไป ดังนั้นเพื่อสามารถเพิ่มประสิทธิภาพให้กับอัลกอริทึมจึงได้ทดลองใช้อัลกอริทึมสกัดคุณลักษณะอื่นร่วมกันทำงาน

2. บทความวิจัยเรื่อง Improved tSNE based manifold dimensional reduction for remote sensing data processing (Song, Wang, Liu, & Choo, 2018) ทำการศึกษาการลดมิติของข้อมูลจากภาพการศึกษาระยะไกล (Remote Sensing) เนื่องจากเทคนิคการลดมิติที่มีอยู่ทำให้ประสิทธิภาพลดลง เพื่อเพิ่มประสิทธิภาพจึงใช้ 2-mixed Gaussian Model (GMM) ในการเพิ่มประสิทธิภาพของอัลกอริทึม tSNE สรุปการกระจายตัวของความน่าจะเป็นแบบใหม่ที่ใช้กับอัลกอริทึม tSNE ช่วยให้การสร้างโครงสร้างในระดับ Local และแยกความแตกต่างระหว่างโครงสร้าง Global ได้อย่างมีนัยสำคัญซึ่งผู้วิจัยได้ศึกษาแนวคิดในการเพิ่มประสิทธิภาพของอัลกอริทึม tSNE ด้วยอัลกอริทึมอื่นจากบทความนี้

3. บทความ Auto-Encoder Based Dimensionality Reduction (Wang, Yao, & Zhao, 2016) งานวิจัยศึกษาการทดลองลดมิติข้อมูลด้วย AUTOENCODER กับวิธีการลดมิติข้อมูลแบบต่าง ๆ คือ PCA, LDA, LLE และ ISOMAP ใช้ชุดข้อมูลที่สร้างขึ้นเองคือชุดข้อมูล 3 มิติที่เป็นโครงสร้างแบบ Circular Helix, Conical Helix และ Planar Helix และชุดข้อมูล MNIST และ Olivetti face การทดลองแบ่งเป็นสองส่วนคือ การประเมินความสามารถในการลดมิติและการหาว่าจำนวน Hidden Layer ส่งผลต่อประสิทธิภาพหรือไม่ สรุปการลดมิติด้วย AUTOENCODER ให้ผลที่แตกต่างอย่างชัดเจนเทียบกับวิธีอื่นคือนอกจากจะลดมิติข้อมูลแล้ว ยังสามารถสกัด

โครงสร้างของข้อมูลออกมาได้และการทดลองจำแนกประเภทชุดข้อมูล MNIST จะให้ผลลัพธ์ที่ดีที่สุดเมื่อจำนวน Hidden Layer ใกล้เคียงหรือเท่ากับมิติของข้อมูล ในบทความนี้ผู้วิจัยได้ทำการศึกษาการใช้ AUTOENCODER เพื่อใช้ในการสกัดคุณลักษณะและพบว่า AUTOENCODER มีประสิทธิภาพมากในการสกัดคุณลักษณะจากคุณลักษณะที่มีอยู่เดิมแต่การไม่สามารถแสดงผลข้อมูลได้เนื่องจากมิติของข้อมูลไม่ได้ถูกลดลงเหลือ 2-3 มิติ จึงได้แนวคิด หากลำอัลกอริทึม AUTOENCODER ทำงานร่วมกับ tSNE อาจเพิ่มประสิทธิภาพให้กับอัลกอริทึมได้อย่างดีรวมไปถึงยังสามารถแสดงข้อมูลได้

4. บทความวิจัยเรื่อง Dimensionality Reduction using PCA and K-Means Clustering for Breast Cancer Prediction (Jamal, Handayani, Septiandri, Ripmiation, & Effendi, 2018) งานวิจัยศึกษาทดลองการจำแนกประเภทของชุดข้อมูล Wisconsin Breast Cancer ใช้อัลกอริทึม Support Vector Machine และ Extreme Gradient Boosting ในการจำแนกประเภทและเปรียบเทียบผล ใช้อัลกอริทึม PCA และ K-Means เพื่อลดมิติและข้อมูลสกัดคุณลักษณะข้อมูล สรุป K-Means ให้ประสิทธิภาพดีกว่าในการลดมิติและนำมาจำแนกประเภทเมื่อเทียบกับ PCA ในบทความนี้ผู้วิจัยได้ศึกษาการใช้อัลกอริทึมสกัดคุณลักษณะและการลดมิติข้อมูลเพื่อเพิ่มประสิทธิภาพให้กับแบบจำลองจำแนกประเภทพบว่าประสิทธิภาพการจำแนกประเภทยังไม่ดีนักจึงได้แนวคิดสมมติฐานหากนำอัลกอริทึมสกัดคุณลักษณะข้อมูลอื่นร่วมทำงานกับอัลกอริทึม tSNE อาจช่วยเพิ่มประสิทธิภาพการจำแนกข้อมูลได้

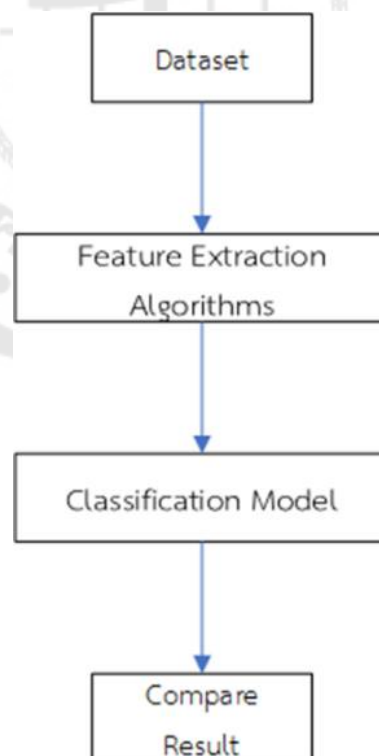
บทที่ 3

วิธีการดำเนินการวิจัย

จากปัญหาที่กล่าวมาในข้างต้น งานวิจัยนี้มุ่งศึกษาผลของการใช้อัลกอริทึม สกัดคุณลักษณะอื่นร่วมกับการใช้อัลกอริทึม tSNE โดยเปรียบเทียบผลของการใช้อัลกอริทึมที่ต่างกันจะมีผลต่อประสิทธิภาพของอัลกอริทึม tSNE อย่างไร บทนี้อธิบายถึงการนำทฤษฎีที่อ้างถึงมาทำการทดลองเพื่อแก้ปัญหาการแสดงผลข้อมูลในมิติสูงโดยใช้การลดมิติข้อมูลเพื่อให้อยู่ในรูปแบบที่สามารถแสดงผลได้และง่ายต่อการเข้าใจ

ในการวิจัยครั้งนี้ ผู้วิจัยได้ดำเนินการตามขั้นตอนดังนี้

1. เลือกชุดข้อมูล
2. ใช้อัลกอริทึมทำการลดมิติข้อมูลและแสดงผลข้อมูลที่ลดมิติแล้ว
3. นำข้อมูลที่ลดมิติแล้วไปทำการวัดประสิทธิภาพโดยใช้อัลกอริทึมจำแนกประเภท
4. เปรียบเทียบและวิเคราะห์ผลลัพธ์ในแต่ละอัลกอริทึม



ภาพประกอบที่ 9 ขั้นตอนการทดลอง

ชุดข้อมูล

Credit Card Fraud Detection เป็นข้อมูลการทำธุรกรรมบัตรเครดิตที่มีข้อมูลการฉ้อโกง เพื่อให้สามารถเรียนรู้ได้ว่าธุรกรรมแบบใดเป็นธุรกรรมที่เป็นการฉ้อโกง ซึ่งบริษัทจะทำการตรวจสอบการทำธุรกรรมนั้นต่อไป ชุดข้อมูลประกอบไปด้วยข้อมูลการทำธุรกรรมจากบัตรเครดิตในเดือนกันยายนปี 2013 โดยผู้ให้บริการบัตรเครดิตในทวีปยุโรป ชุดข้อมูลนี้แสดงข้อมูลธุรกรรมที่เกิดขึ้นจำนวน 2 วัน ซึ่งพบว่ามีข้อมูลธุรกรรมที่มีการฉ้อโกงจำนวน 492 รายการจากการทำธุรกรรมทั้งหมด 284087 รายการ ชุดข้อมูลนี้มีจำนวนข้อมูลธุรกรรมที่มีการฉ้อโกงกับจำนวนข้อมูลธุรกรรมปกติต่างกันอย่างมากคือมีข้อมูลธุรกรรมที่เป็นการฉ้อโกงเพียง 0.172% ชุดข้อมูลเป็นแบบตัวเลขซึ่งเป็นผลลัพธ์ของการทำ PCA ซึ่งทำให้ไม่สามารถบอกได้ว่าเป็นข้อมูลดั้งเดิมเป็นข้อมูลใด มีจำนวนคุณลักษณะทั้งหมดจำนวน 28 คุณลักษณะ คุณลักษณะที่ไม่ได้ผ่านการทำ PCA คือ 'time' และ 'amount' ส่วนคุณลักษณะ 'class' จะเป็นการแทนข้อมูลธุรกรรมที่มีการฉ้อโกงด้วยเลข 1 และข้อมูลธุรกรรมปกติด้วยเลข 0 (Bruxelles), 2018)

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20	V21	V22	V23	V24	V25	V26	V27	V28	Amount	Class
0	-1.35981	-0.07278	2.536347	1.378155	-0.33832	0.462388	0.239599	0.098988	0.363787	0.090794	-0.5516	-0.6178	-0.99139	-0.31177	1.468177	-0.47004	0.207971	0.025791	0.403993	0.253412	-0.01831	0.277838	-0.11047	0.066928	0.128339	-0.18911	0.133558	-0.02105	149.62	0
0	1.191857	0.266151	0.16648	0.448154	0.060018	-0.08236	-0.07788	0.085102	-0.25543	-0.16697	1.612727	1.065235	0.489095	-0.14377	0.633558	0.463917	-0.1148	-0.18336	-0.14578	-0.06908	-0.22578	-0.63867	0.101288	-0.33985	0.16717	0.121895	-0.00898	0.014724	2.69	0
1	-1.35835	-1.34016	1.773209	0.37978	-0.3032	1.804599	0.791461	0.247678	-1.51465	0.207643	0.624501	0.066084	0.717279	-0.16395	0.345485	-2.00695	1.109969	-0.12136	-0.26186	0.247998	0.716739	0.908412	-0.68828	-0.32784	-0.1391	0.05335	-0.05975	378.66	0	
1	0.96427	-0.18521	1.792953	0.86329	-0.01031	1.247203	0.237609	0.377436	-1.88702	-0.05495	0.23649	0.178238	0.50737	-0.28792	-0.61542	-1.05965	-0.68409	1.965775	-1.23262	-0.20804	-0.10801	0.052574	-0.18932	-1.17558	0.647176	-0.21189	0.062723	0.061458	129.5	0
2	-1.15823	0.877737	1.548718	0.400304	-0.40719	0.095921	0.592941	-0.27053	0.817739	0.753074	-0.82284	0.538136	1.345852	-1.11987	0.171211	-0.45145	-0.23703	-0.03819	0.803487	0.408542	-0.09043	0.11746	0.141287	-0.20601	0.502292	0.219422	0.211533	69.99	0	
2	0.42597	0.960523	1.141109	-0.16825	0.420987	-0.02973	0.476201	0.260314	-0.58867	-0.37411	1.341262	0.359894	-0.35809	-0.13713	0.515717	0.401726	-0.05813	0.068653	-0.03319	0.084968	-0.26825	-0.55982	-0.0264	-0.37143	-0.23279	0.105915	0.253844	0.08108	3.67	0
4	2.20658	0.141004	0.045371	0.208513	0.103981	0.272708	-0.005158	0.061123	0.464996	-0.05922	-1.41691	-0.15383	-0.75108	0.867972	0.050244	-0.44309	0.002821	-0.61199	-0.04538	0.21963	-0.16772	0.27071	-0.1541	-0.78006	0.750137	-0.25724	0.034507	0.005368	4.99	0
7	0.64427	1.417864	1.07438	-0.4922	0.948954	0.426118	1.120631	-3.80786	0.615375	1.248978	-0.62947	0.295474	1.757964	-1.32387	0.686133	-0.07613	-1.22213	-0.35822	0.324505	-0.15674	1.943455	-0.15445	0.057504	-0.64971	-0.41527	-0.05163	-1.20692	-1.08534	40.8	0
7	-0.89429	0.288157	-0.11319	-0.27153	2.689999	3.721818	0.370145	0.851084	-0.39205	-0.41043	-0.70512	-0.11045	-0.28625	0.074355	-0.33878	-0.21008	-0.49977	0.118765	0.570328	0.052736	-0.07343	-0.26809	-0.20423	0.101592	0.373205	-0.38416	0.011747	0.142404	93.2	0
9	-0.33826	1.115993	1.044367	-0.22219	0.499961	-0.24676	0.651583	0.069539	-0.79873	-0.36685	1.017614	0.83639	1.006841	-0.44352	0.150219	0.739453	-0.54098	0.476677	0.451773	0.202711	-0.14691	-0.63375	-0.12079	-0.38055	-0.06973	0.094199	0.246219	0.083076	3.68	0
10	1.448044	-1.17634	0.91386	1.37567	-1.97138	-0.62615	-1.43234	0.048436	1.72041	1.626659	1.199644	0.67144	-0.13397	-0.09055	0.23093	0.019187	0.253415	0.854344	-0.22137	-0.36723	-0.00901	0.31894	0.02774	0.00512	0.251467	-0.12948	0.04265	0.010253	7.8	0
10	0.984978	0.631509	-0.8743	-0.95402	2.624584	3.317027	0.470455	0.582427	-0.55889	0.309753	-0.25912	-0.32614	-0.09005	0.362832	0.928094	-0.12949	-0.80998	0.359985	0.707654	0.125962	0.049924	0.288422	0.00913	0.96731	-0.76731	-0.49221	0.962472	0.05434	9.99	0
12	1.249999	-1.21264	0.88393	-1.2349	-1.48542	-0.75323	-0.6894	-0.22749	-0.29401	1.523729	0.227866	-0.24268	1.205417	-0.31763	0.723675	-0.81561	0.873936	-0.84779	-0.68319	-0.10276	-0.23181	-0.48329	0.084668	0.392831	0.161135	-0.35499	0.026416	0.042422	121.5	0
12	0.28772	0.282813	2.71252	-0.1784	0.397544	-0.09672	0.115982	-0.22108	0.46023	-0.77866	0.523287	-0.01108	-0.17849	-0.65556	-0.19999	0.124005	-0.8605	-0.98292	-0.1532	-0.05688	0.074412	-0.17141	0.104744	0.548265	0.104094	0.021491	0.021293	27.5	0	
12	-2.78185	-0.27277	1.64175	1.767473	-0.13699	0.807196	-0.42391	-1.90711	0.705713	1.151087	0.848355	0.792844	0.370448	0.73498	-0.409796	-0.30306	-0.13587	0.778261	2.213886	1.58212	1.151661	0.22182	1.029566	0.02817	0.23275	-0.23564	-0.16478	0.09015	5.8	0
12	-0.75242	0.345485	2.057323	-1.46864	-1.15839	-0.07785	-0.06858	0.005603	-0.43617	0.747731	-0.78398	-0.77041	1.047627	-1.0666	1.109953	1.660134	-0.27927	-0.41999	0.432335	0.265451	0.049625	1.55365	-0.25657	-0.06508	-0.05912	-0.08709	-0.181	0.129394	15.99	0
12	1.035215	-0.0401	1.267332	1.289091	-0.736	0.288069	-0.58806	0.18938	0.782333	-0.26798	-0.45031	0.896708	0.70388	-0.46865	0.354574	-0.24663	-0.00921	-0.59591	-0.57568	-0.11391	-0.02461	0.196002	0.013078	0.364298	-0.38226	0.092809	0.037051	12.99	0	
13	-0.43691	0.918966	0.24591	-0.72722	0.915679	-0.12787	0.707842	0.087962	-0.68527	-0.73798	0.124098	0.277192	0.252624	-0.2919	-0.18432	1.143174	-0.92871	0.68047	0.025436	-0.04702	-0.1948	-0.67264	-0.11686	-0.88839	-0.34241	-0.04903	0.079692	0.131024	0.89	0
14	-0.40126	-5.45031	1.188055	1.786239	3.091026	-1.76541	-1.53974	0.160842	1.230309	0.345177	0.91723	0.970117	-0.26857	-0.47913	-0.52661	0.472008	-0.72348	0.070261	-0.40687	-0.19885	-0.5036	0.98446	-0.458589	0.041219	-0.48163	-0.62127	0.392025	0.949094	46.8	0
15	1.492396	-1.02935	0.545795	-1.43803	-1.55543	-0.72096	-1.08066	-0.05313	-1.87688	1.638076	1.077542	-0.63205	-0.41696	0.052011	-0.04298	-0.16681	0.304241	0.554432	0.05423	-0.38791	-0.17765	-0.17507	0.040002	0.295814	-0.22038	0.022298	0.007602	5	0	
16	0.694885	-1.36182	1.029221	0.834159	-1.19321	1.309109	-0.87859	0.44539	-0.4462	0.568521	1.019151	1.298329	0.42048	-0.37265	-0.80718	-0.04436	0.153663	0.625847	-1.30041	-0.13833	-0.19558	-0.57196	-0.05088	-0.42223	0.088533	0.063499	231.71	0		
17	0.962498	0.328461	-0.17148	1.209204	1.129586	1.696038	0.107712	0.521502	-1.19331	0.724396	1.69053	0.406774	-0.95842	0.983739	0.720911	-0.20223	0.402484	-1.73716	-0.02761	-0.26952	0.149997	0.046851	-1.57187	0.990814	0.199964	0.016371	-0.01461	34.09	0	
18	1.16616	0.90212	-0.0673	2.815169	0.428804	0.084714	0.24147	0.138062	-0.98916	0.921275	0.743786	-0.53138	-2.10531	1.24687	0.000793	0.042425	-0.45448	-0.09887	-0.8166	-0.30717	0.058702	-0.06037	-0.10183	-0.7042	0.6012	0.109556	-0.04052	0.01142	2.38	0
18	0.247491	0.277666	1.185471	-0.0926	-1.31439	-0.15012	-0.94836	-1.61794	1.544071	-0.82988	-0.5832	0.524933	-0.45338	0.081393	1.553204	-1.39689	0.783131	0.436621	2.177807	-0.23098	1.60518	0.200454	-0.18355	0.472073	0.820991	-0.24783	0.106634	0.250475	22.75	0
22	-1.94653	-0.0449	-0.40937	-1.01306	2.941988	2.955053	-0.08306	0.855546	0.049967	0.573743	-0.08126	-0.21575	0.044161	0.03908	1.907718	0.578843	-0.97567	0.044063	0.888603	-0.22672	-0.57993	-0.79923	0.8703	0.983421	0.321201	0.14905	0.707519	0.0146	0.89	0
22	-0.27029	-0.12146	1.22021	0.410008	0.265188	-0.99554	0.545985	-0.10463	0.478664	0.149451	-0.85857	-0.18052	-0.85523	-0.2779	-0.21187	-0.3332	0.010751	-0.48847	0.505751	-0.38669	-0.40364	-0.2274	0.742435	0.398531	0.249212	0.274026	0.399969	0.243232	26.45	0
23	1.173285	0.93488	0.383905	1.131683	-0.17258	-0.91405	0.369025	-0.12728	-0.24465	-0.04624	-0.14342	0.97935	1.492283	0.104118	0.763478	-0.01458	-0.51164	-0.31506	-0.39009	0.027878	0.067003	0.279812	-0.15049	0.455041	0.748232	-0.19708	0.016388	0.030941	41.88	0
23	1.322707	-1.07404	0.345555	0.570638	-0.38676	-0.83108	-0.2649	-0.22098	-1.07142	0.868559	-0.64151	-0.11132	0.361485	0.171945	0.782167	-1.35587	-0.21694	1.271765	-1.24062	-0.52295	-0.28438	-0.32336	-0.05771	0.347153	0.559639	-0.28016	0.042335	0.028822	16	0
23	-0.41429	0.905437	1.27433	1.473471	0.007443	-0.20033	0.740228	-0.02925	-0.59339	-0.34619	-0.01214	0.786796	0.639594	-0.08632	0.078084	-1.40592	0.737992	-0.94289	0.343969	0.097008	0.077327	0.457311	-0.10485	0.642522	-0.18389	-0.27746	0.142687	0.152665	35	0
23	1.059387	-0.17532	1.26613	1.86161	-0.786	0.578455	-0.78708	0.401046	0.8995	-0.06474	1.048202	1.005618	0.524	-0.03991	-0.21868	-0.004476	-0.19355	0.042388	-0.27783	-0.17802	0.123676	0.133734	0.014462	0.020951	0.294368	-0.39507	0.081461	0.012422	12.99	0
24	1.237429	0.061045	0.380526	0.761564	-0.39377	-0.49408	0.006494	-0.13386	0.43881	-0.20776	-0.92918	0.527106	0.346876	-0.10534	-0.11858	-0.63379	0.548416	-0.06635	-0.26585	-0.55309	-0.04647	0.079168	0.509136	0.288858	-0.0227	0.011836	17.28	0		
25	1.114009	0.085546	0.493702	1.33576	-0.30019	-0.01075	-0.11876	0.188617	0.205687	0.082262	1.133556	0.626699	-1.49278	0.520788	-0.674595	-0.52911	0.158256	-0.39675	-0.14571	-0.27383	-0.08325	-0.00476	-0.03147	-0.03147	-0.03147	-0.03147	-0.03147	-0.03147	17	

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fractal_dimension_mean	radius_se	texture_se	perimeter_se	area_se	smoothness_se	compactness_se	cc
842302 M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095	0.9053	8.589	153.4	0.006399	0.04904		
842517 M	20.57	17.77	132.9	1326	0.09474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.7339	3.398	74.08	0.002225	0.01308		
84300903 M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	0.7456	0.7869	4.585	94.03	0.00615	0.04006		
84348301 M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	0.4956	1.156	3.445	27.23	0.00911	0.07458		
84358402 M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	0.7572	0.7813	5.438	94.44	0.01149	0.02461		
843786 M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	0.3345	0.8902	2.217	27.19	0.00751	0.03345		
844259 M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	0.4467	0.7732	3.18	53.91	0.004314	0.01382		
84458202 M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	0.5835	1.377	3.856	50.96	0.00805	0.03029		
844981 M	13	21.82	87.5	519.8	0.1273	0.1932	0.1839	0.09353	0.235	0.07389	0.3063	1.002	2.406	24.32	0.005731	0.03502		
84501001 M	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243	0.2976	1.599	2.039	23.94	0.007149	0.07217		
845636 M	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697	0.3795	1.187	2.466	40.51	0.004029	0.009269		
84610002 M	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06696	0.1842	0.06082	0.5056	0.9849	3.564	54.16	0.005771	0.04061		
846236 M	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078	0.9555	3.568	11.07	116.2	0.003139	0.08297		
846381 M	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.05338	0.4033	1.078	2.903	36.58	0.009769	0.03126		
84667401 M	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.07682	0.2121	1.169	2.061	19.21	0.006429	0.05936		
84799002 M	14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303	0.07077	0.37	1.033	2.879	32.55	0.005607	0.0424		
848406 M	14.48	20.13	94.74	684.5	0.09867	0.072	0.07395	0.05259	0.1586	0.05922	0.4727	1.24	3.195	45.4	0.005718	0.01162		
84862001 M	16.13	20.68	108.1	798.8	0.117	0.2022	0.1722	0.1028	0.2164	0.07356	0.5692	1.073	3.854	54.18	0.007026	0.02501		
849014 M	19.81	22.15	130	1280	0.09831	0.1027	0.1479	0.09498	0.1582	0.05395	0.7582	1.017	5.865	112.4	0.006494	0.01893		
8510426 B	13.54	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	0.1885	0.05766	0.2699	0.7886	2.058	23.56	0.008462	0.0146		
8510653 B	13.08	15.71	85.63	520	0.1075	0.127	0.04588	0.0311	0.1987	0.06811	0.1852	0.7477	1.383	14.67	0.004977	0.01898		
8510824 B	9.504	12.44	60.34	273.9	0.1024	0.06492	0.02956	0.02078	0.1815	0.05905	0.2773	0.9768	1.909	15.7	0.009606	0.01412		
8511133 M	15.34	14.26	102.5	704.4	0.1073	0.2135	0.2077	0.09756	0.2521	0.07032	0.4188	0.7096	3.384	44.91	0.006789	0.05328		
8511409 M	21.16	21.04	137.2	1404	0.09426	0.1027	0.1067	0.08637	0.1769	0.05736	0.6917	1.177	4.303	93.99	0.004728	0.01759		

ภาพประกอบที่ 11 แสดงตัวอย่างชุดข้อมูล Breast Cancer Wisconsin Data

Heart Disease UCI เป็นชุดข้อมูลโรคหัวใจประกอบด้วย 14 คุณลักษณะเช่นอายุ เพศ ประเภทการเจ็บหน้าอก ความดันโลหิต ปริมาณคอเลสเตอรอลในกระแสเลือด ปริมาณน้ำตาลในกระแสเลือด ผลการวัดคลื่นหัวใจ จำนวนคลาสที่เป็นโรคหัวใจ 165 แถว ไม่เป็นโรคหัวใจ 138 แถว (Andras Janosi, 1988)

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
44	1	1	120	263	0	1	173	0	0	2	0	3	1
52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
58	0	3	150	283	1	0	162	0	1	2	0	2	1
50	0	2	120	219	0	1	158	0	1.6	1	0	2	1
58	0	2	120	340	0	1	172	0	0	2	0	2	1
66	0	3	150	226	0	1	114	0	2.6	0	0	2	1
43	1	0	150	247	0	1	171	0	1.5	2	0	2	1
69	0	3	140	239	0	1	151	0	1.8	2	2	2	1
59	1	0	135	234	0	1	161	0	0.5	1	0	3	1
44	1	2	130	233	0	1	179	1	0.4	2	0	2	1
42	1	0	140	226	0	1	178	0	0	2	0	2	1
61	1	2	150	243	1	1	137	1	1	1	0	2	1

ภาพประกอบที่ 12 แสดงตัวอย่างชุดข้อมูล Heart Disease UCI

กำหนดอัลกอริทึมที่ใช้ทำการสกัดคุณลักษณะ

ในงานวิจัยนี้ผู้วิจัยได้กำหนดอัลกอริทึมสกัดคุณลักษณะ PCA , Kernel PCA , AUTOENCODER และ tSNE เพื่อใช้ในการสกัดคุณลักษณะและลดมิติข้อมูล

กำหนดอัลกอริทึมที่ใช้จำแนกข้อมูล

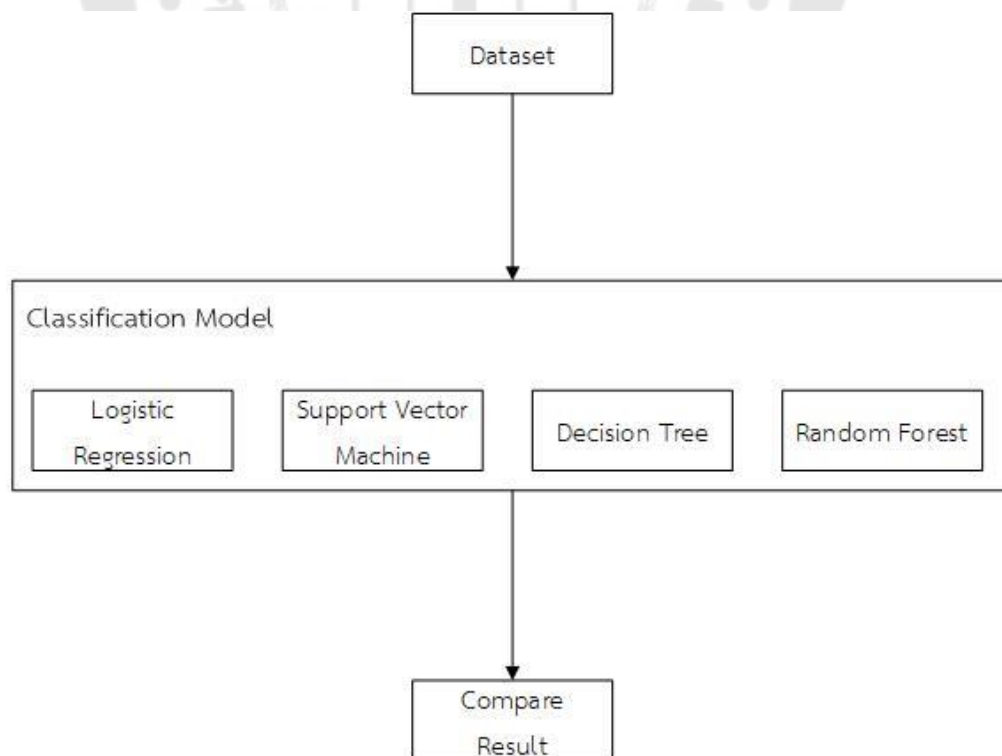
ในงานวิจัยนี้ผู้วิจัยได้กำหนดอัลกอริทึมสำหรับจำแนกประเภทข้อมูล Logistic Regression , Support Vector Machine , Decision Tree และ Random Forest

วิธีการดำเนินงานวิจัย

งานวิจัยนี้จะทำการทดลองใช้อัลกอริทึมสกัดคุณลักษณะจากชุดข้อมูลทั้ง 3 ชุดโดยแบ่งการทดลองออกเป็น 8 กลุ่มการทดลอง ได้แก่

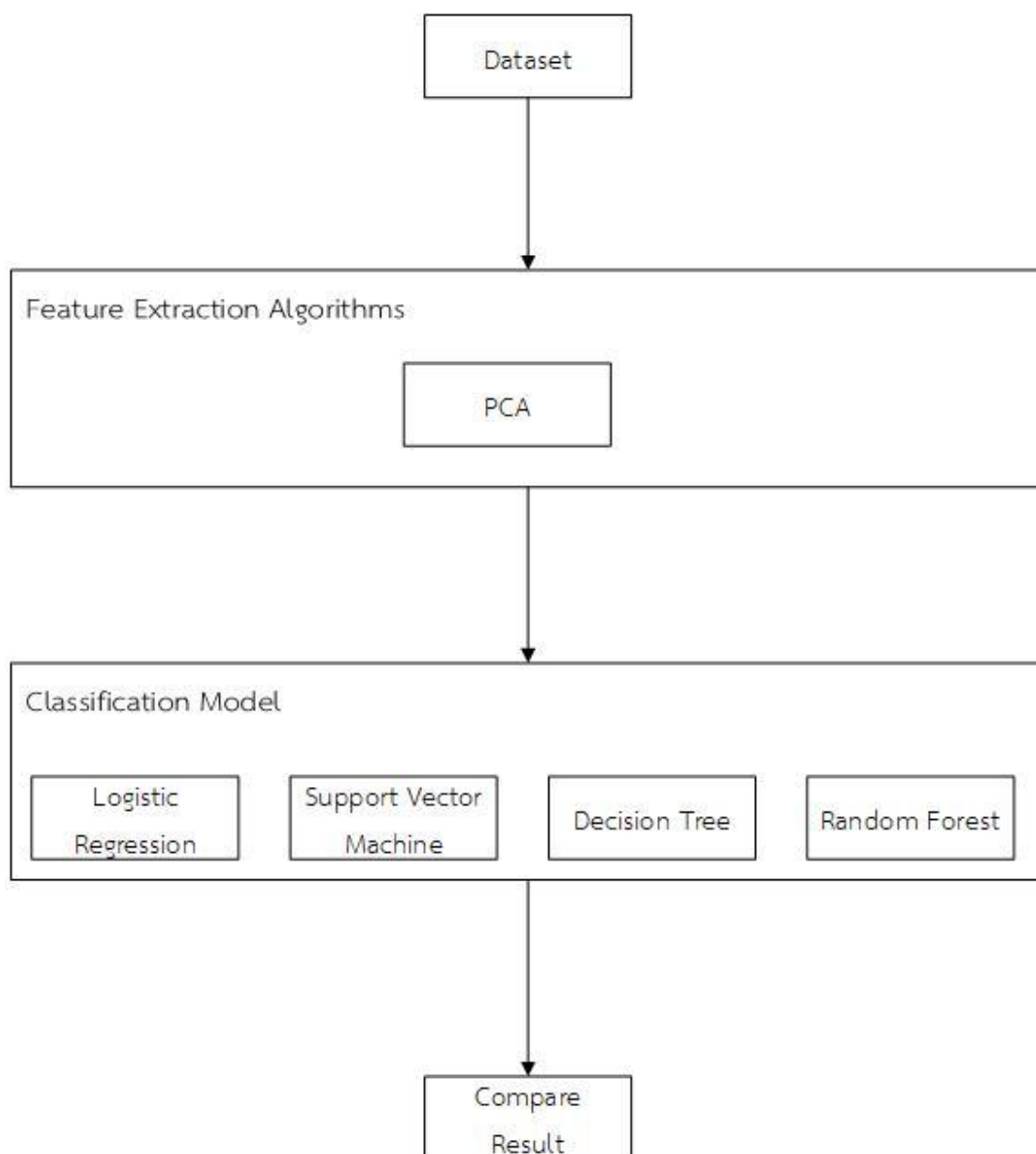
1. การทดลองโดยไม่ผ่านการใช้อัลกอริทึมสกัดคุณลักษณะใดเพื่อใช้ผลลัพธ์ของการจำแนกประเภทเป็นบรรทัดฐาน โดยในการทดลองนี้ จะใช้เป็นการทดลองฐาน (Baseline) เพื่อดูว่าหากไม่มีการลดมิติเลยจะมีประสิทธิภาพในการจำแนกข้อมูลเพียงใด ดังแสดงในภาพประกอบที่

13



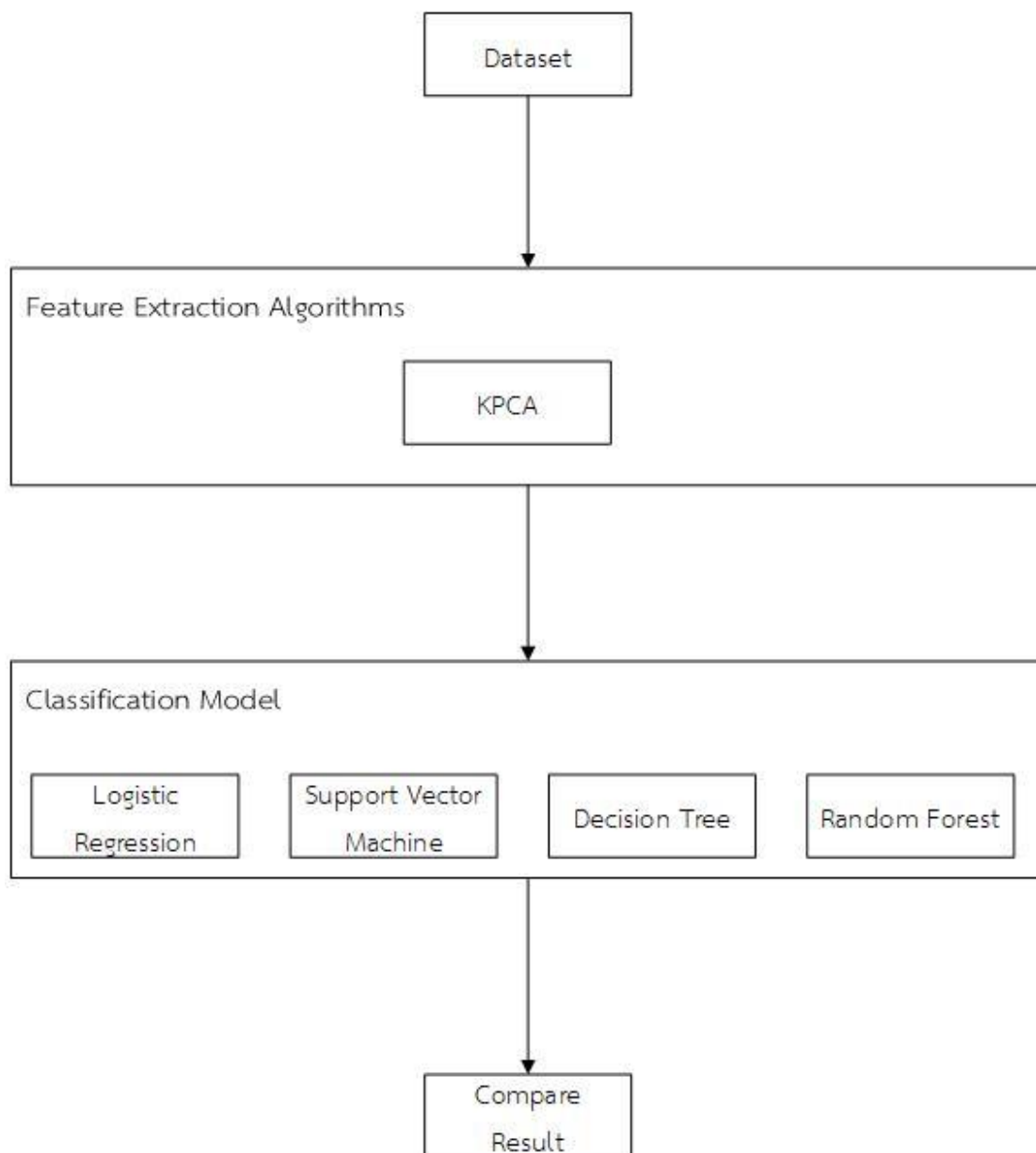
ภาพประกอบที่ 13 แสดงขั้นตอนการทดลองโดยไม่ใช้อัลกอริทึมสกัดคุณลักษณะ

2. การทดลองถัดมาใช้อัลกอริทึม PCA และทำการทดลองจำแนกประเภทเพื่อวัดประสิทธิภาพการทดลองนี้ มีจุดมุ่งหมายคือต้องการศึกษาว่าหากใช้อัลกอริทึม PCA แต่เพียงอย่างเดียวจะมีประสิทธิภาพในการจำแนกข้อมูลเพียงใด เพื่อเปรียบเทียบกับกรณีที่ใช้ทั้งอัลกอริทึม PCA ร่วมกับ อัลกอริทึม tSNE ดังแสดงในภาพประกอบที่ 14



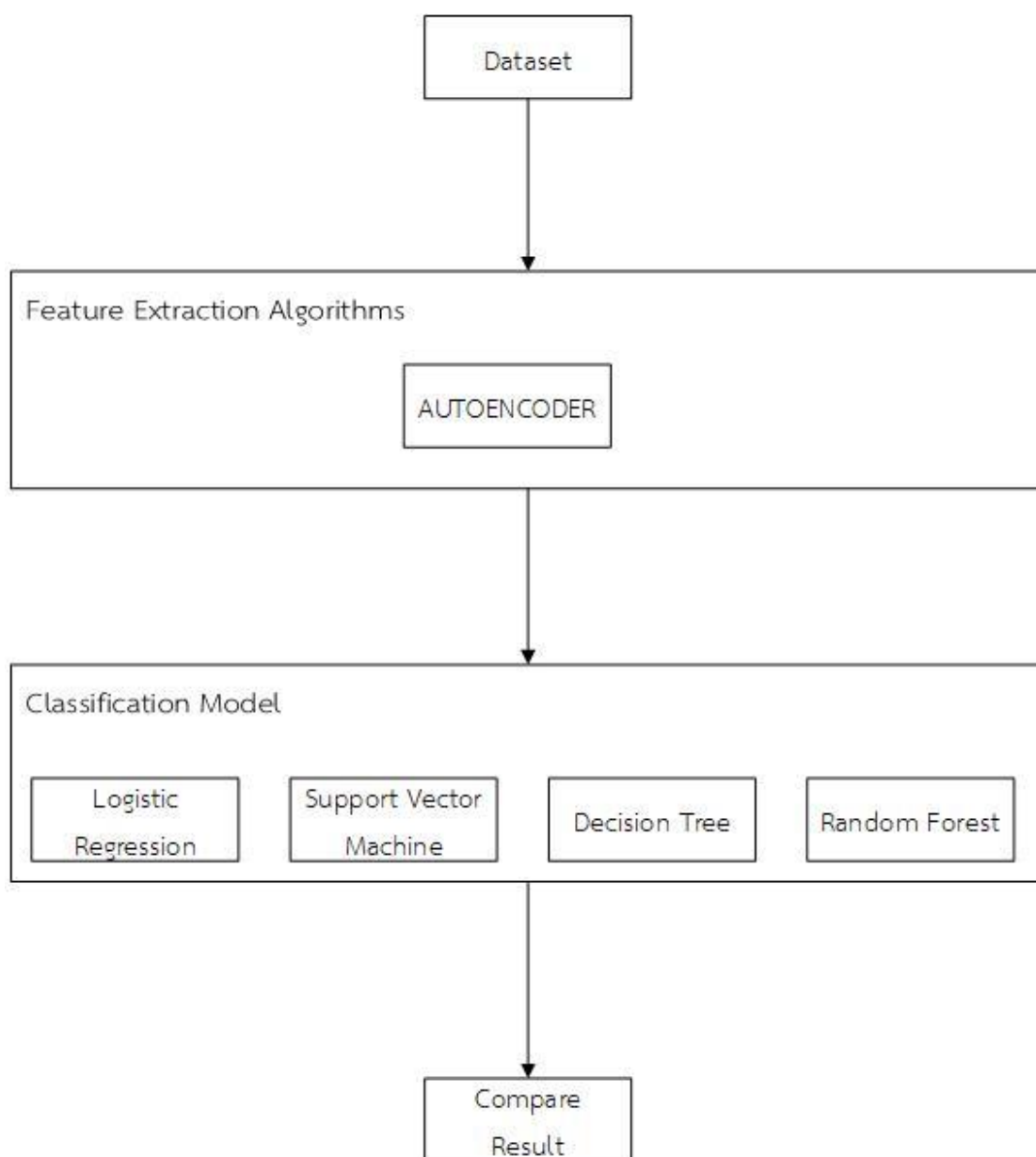
ภาพประกอบที่ 14 แสดงขั้นตอนการทดลองโดยใช้อัลกอริทึมสกัดคุณลักษณะ PCA

3. การทดลองโดยใช้อัลกอริทึม Kernel PCA และทำการทดลองจำแนกประเภท เพื่อวัดประสิทธิภาพ การทดลองนี้มีจุดมุ่งหมายคือต้องการศึกษาว่า หากใช้อัลกอริทึม Kernel PCA แต่เพียงอย่างเดียวจะมีประสิทธิภาพในการจำแนกข้อมูลเพียงใด เพื่อเปรียบเทียบกับกรณีที่ใช้ทั้งอัลกอริทึม Kernel PCA ร่วมกับ อัลกอริทึม tSNE ดังแสดงในภาพประกอบที่ 15



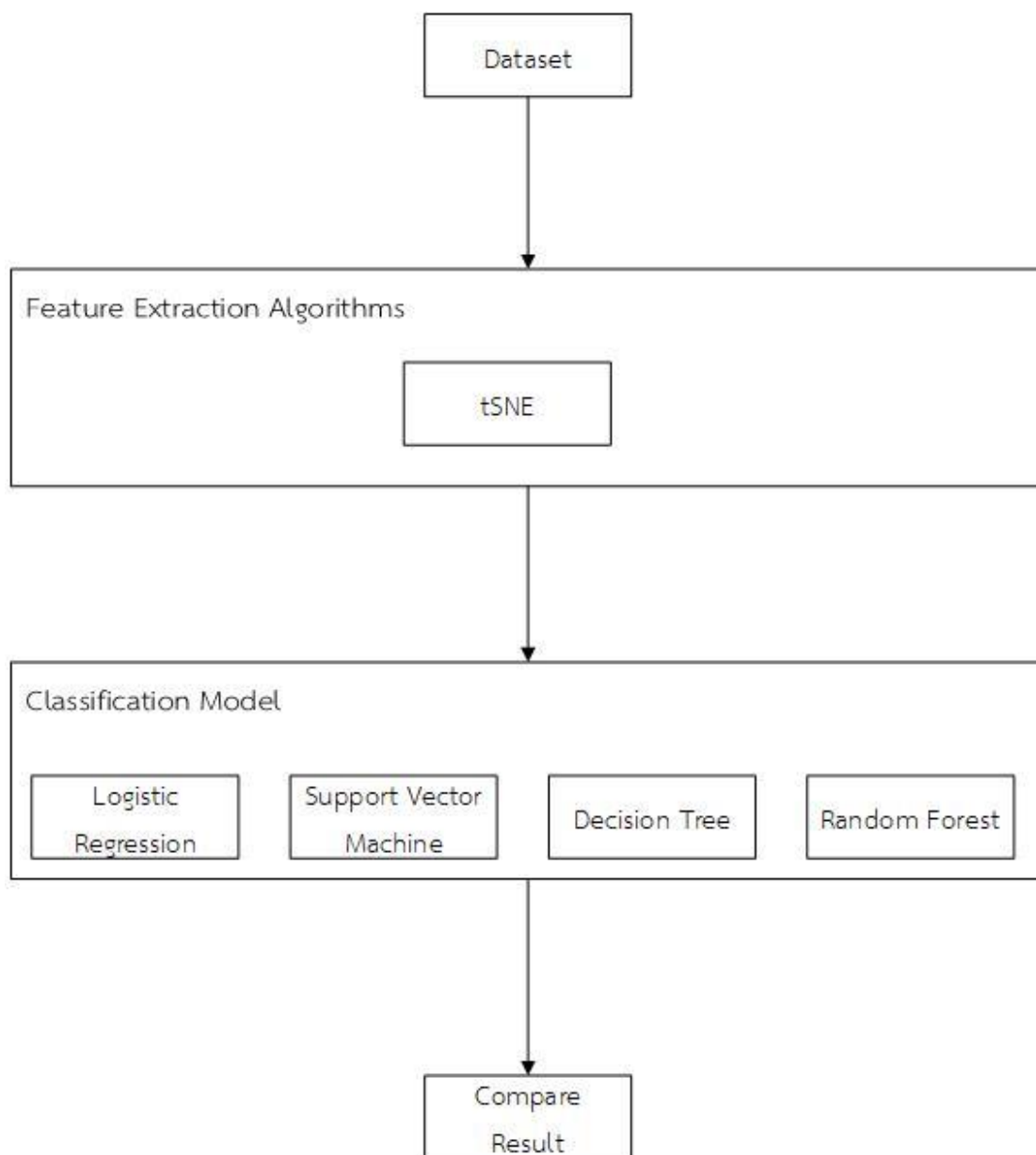
ภาพประกอบที่ 15 แสดงขั้นตอนการทดลองโดยใช้อัลกอริทึมสกัดคุณลักษณะ KPCA

4. การทดลองโดยใช้อัลกอริทึม AUTOENCODER และทำการทดลองจำแนกประเภทเพื่อวัดประสิทธิภาพ การทดลองนี้ มีจุดมุ่งหมายคือ ต้องการศึกษาว่า หากใช้อัลกอริทึม AUTOENCODER แต่เพียงอย่างเดียว จะมีประสิทธิภาพในการจำแนกข้อมูลเพียงใด เพื่อเปรียบเทียบกับกรณีที่ใช้ทั้งอัลกอริทึม AUTOENCODER ร่วมกับอัลกอริทึม tSNE ดังแสดงในภาพประกอบที่ 16



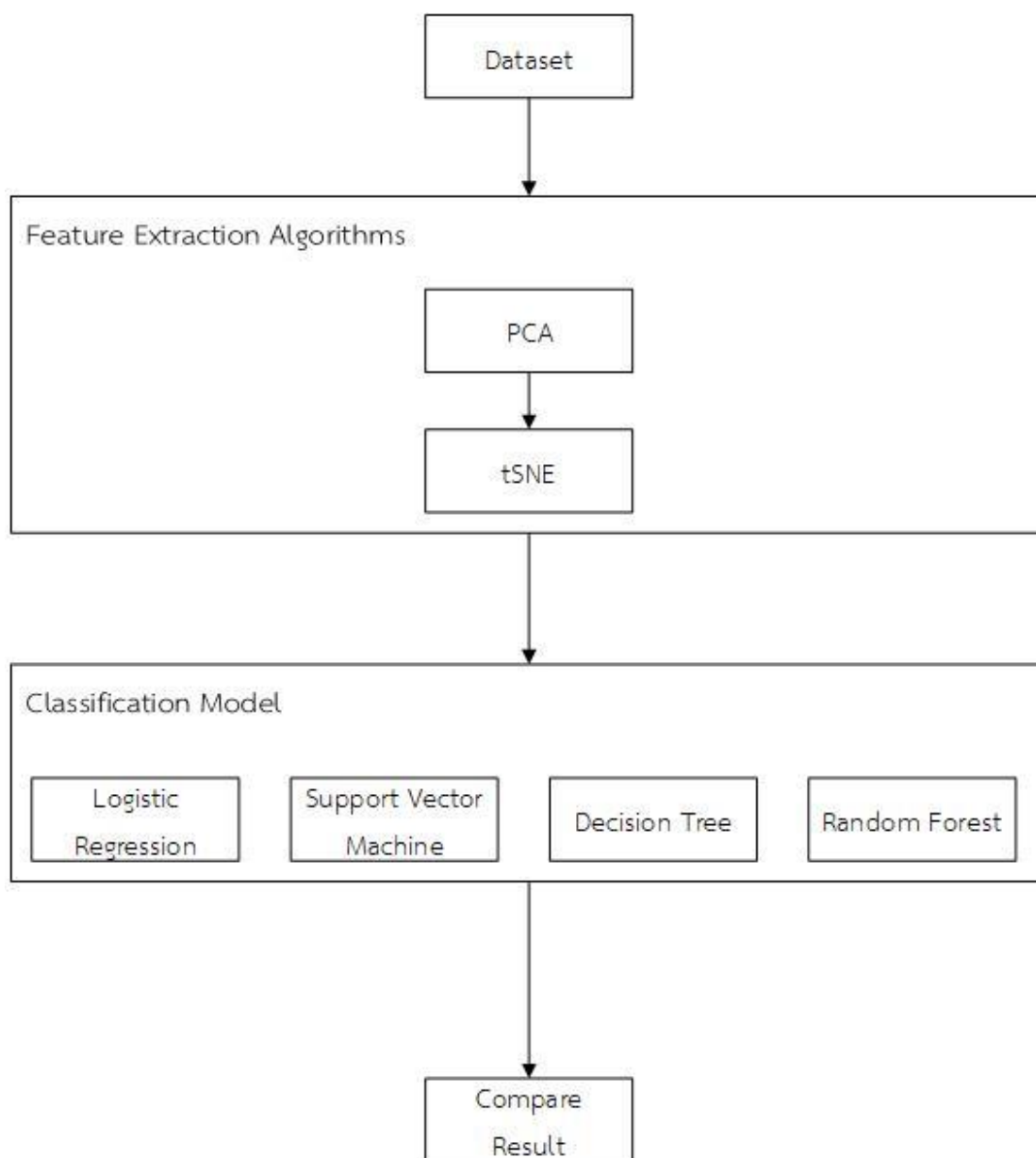
ภาพประกอบที่ 16 แสดงขั้นตอนการทดลองโดยใช้อัลกอริทึมสกัดคุณลักษณะ AUTOENCODER

5. การทดลองโดยใช้อัลกอริทึม tSNE และทำการทดลองจำแนกประเภทเพื่อวัดประสิทธิภาพดังแสดงในภาพประกอบที่ 17



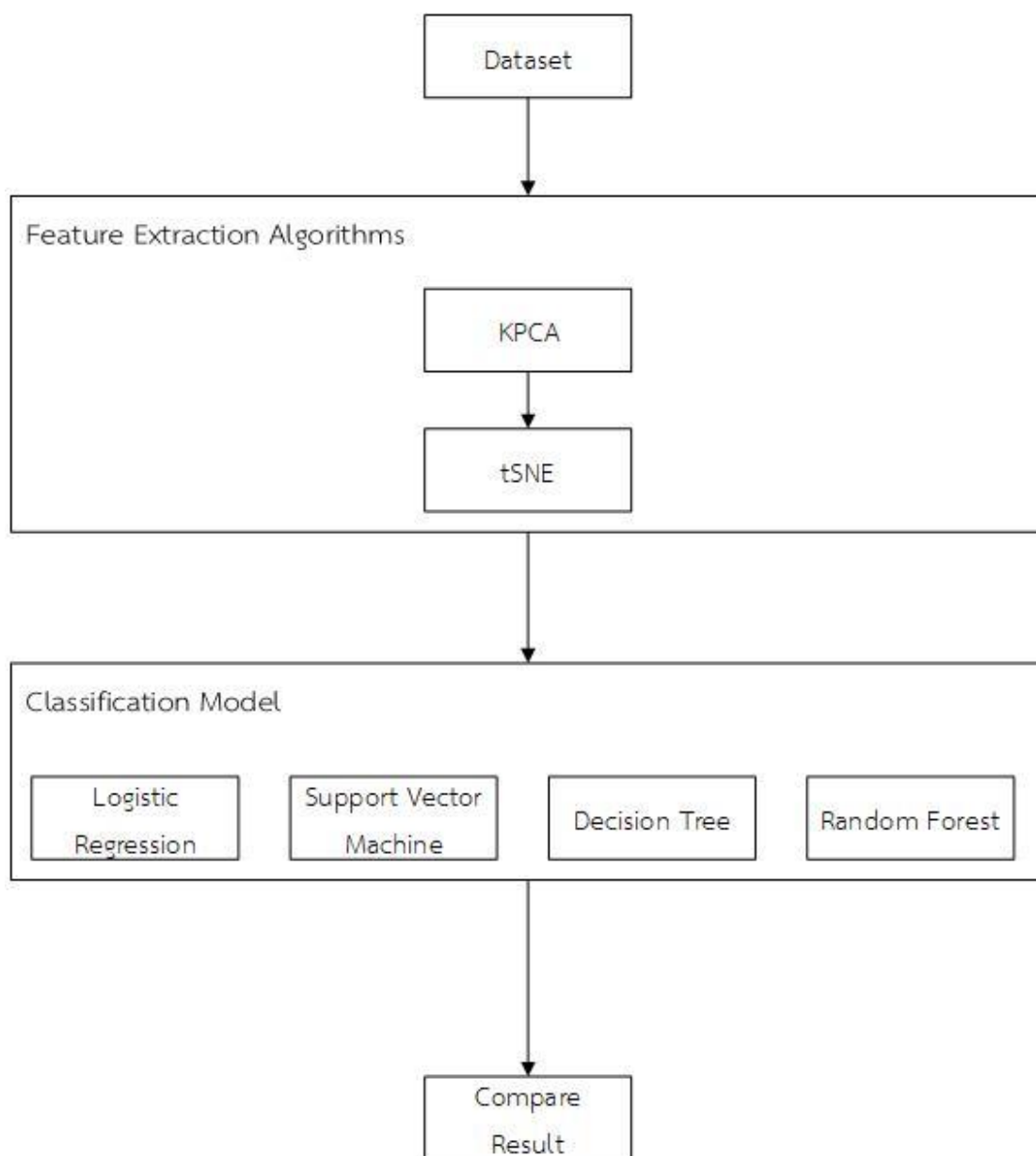
ภาพประกอบที่ 17 แสดงขั้นตอนการทดลองโดยใช้อัลกอริทึมสกัดคุณลักษณะ tSNE

6. การทดลองโดยใช้อัลกอริทึม PCA ร่วมกับ tSNE และทำการทดลองจำแนกประเภทเพื่อวัดประสิทธิภาพดังแสดงในภาพประกอบที่ 18



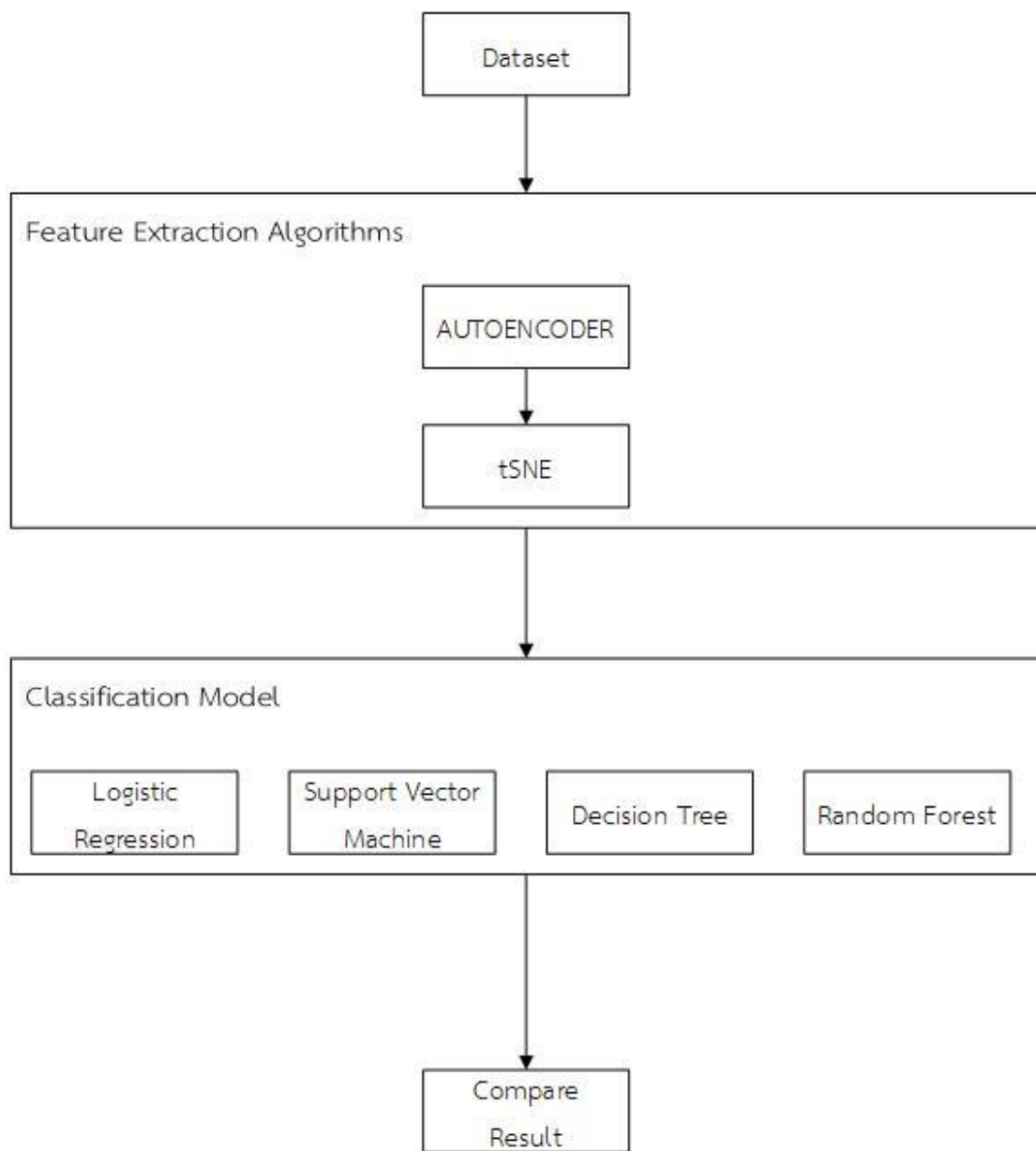
ภาพประกอบที่ 18 แสดงขั้นตอนการทดลองโดยใช้อัลกอริทึมสกัดคุณลักษณะ PCA ร่วมกับ tSNE

7. การทดลองโดยใช้อัลกอริทึม Kernel PCA ร่วมกับ tSNE และทำการทดลองจำแนกประเภทเพื่อวัดประสิทธิภาพดังแสดงในภาพประกอบที่ 19



ภาพประกอบที่ 19 แสดงขั้นตอนการทดลองโดยใช้อัลกอริทึมสกัดคุณลักษณะ KPCA ร่วมกับ tSNE

8. การทดลองโดยใช้อัลกอริทึม AUTOENCODER ร่วมกับ tSNE และทำการทดลองจำแนกประเภทเพื่อวัดประสิทธิภาพดังแสดงในภาพประกอบที่ 20



ภาพประกอบที่ 20 แสดงขั้นตอนการทดลองโดยใช้อัลกอริทึมสกัดคุณลักษณะ AUTOENCODER ร่วมกับ tSNE

จากนั้นจึงทำเปรียบเทียบประสิทธิภาพที่ได้จากการทดลองชุดต่าง ๆ เพื่อศึกษาและเลือกอัลกอริทึมสกัดคุณลักษณะที่เมื่อนำมาใช้งานร่วมกับอัลกอริทึม tSNE แล้วทำให้ข้อมูลจากการลดมิติมีประสิทธิภาพดีที่สุดเมื่อใช้งานในแบบจำลองแยกประเภท

บทที่ 4

ผลการศึกษา

ในการวิจัยการศึกษาผลของการใช้อัลกอริทึมสกัดคุณลักษณะต่ออัลกอริทึม tSNE ผู้วิจัยได้ดำเนินการวิจัยโดยศึกษาตามขั้นตอนต่าง ๆ ตลอดจนวัดประสิทธิภาพ เพื่อให้บรรลุจุดประสงค์ที่กำหนดไว้ได้ดังนี้

1. ผลลัพธ์การเปรียบเทียบประสิทธิภาพของแบบจำลองของชุดข้อมูล Credit Card Fraud Detection
2. ผลลัพธ์การเปรียบเทียบประสิทธิภาพของแบบจำลองของชุดข้อมูล Breast Cancer
3. ผลลัพธ์การเปรียบเทียบประสิทธิภาพของแบบจำลองของชุดข้อมูล Heart Disease

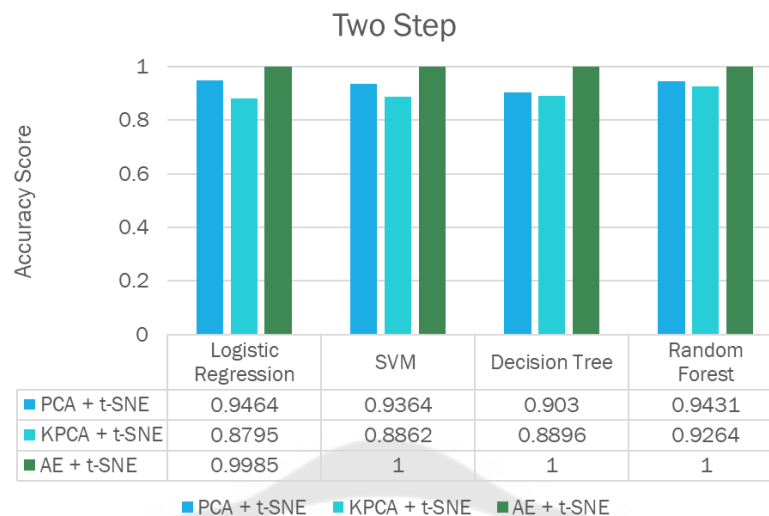
ผลลัพธ์การเปรียบเทียบประสิทธิภาพของแบบจำลองของชุดข้อมูล Credit Card Fraud Detection

จากการทดลองเพื่อทำนายว่าเป็นข้อมูลการทำรายการที่เป็นการโกงหรือไม่ โดยใช้แบบจำลอง Logistic Regression, Support Vector Machine, Decision Tree และ Random Forest ได้ผลลัพธ์ดังนี้



ภาพประกอบที่ 21 แสดงผลการทำนายของชุดข้อมูล Credit Card Fraud Detection โดยใช้
อัลกอริทึมสกัดคุณลักษณะเดียว

จากรูปภาพที่ 10 แสดงผลการทำนายของชุดข้อมูล Credit Card Fraud Detection โดยใช้
อัลกอริทึมสกัดคุณลักษณะเดียว พบว่าในผลการทำนายของแบบจำลอง Logistic Regression
ได้ค่า accuracy ในกรณีไม่ใช้อัลกอริทึมสกัดคุณลักษณะ 0.9598 ใช้อัลกอริทึม PCA 0.9364 ใช้อัลกอริทึม
Kernel PCA 0.8929 ใช้อัลกอริทึม AUTOENCODER 1 ใช้อัลกอริทึม tSNE 0.7993 ผลการทำนายของแบบจำลอง
Support Vector Machine ได้ค่า accuracy ในกรณีไม่ใช้อัลกอริทึมสกัดคุณลักษณะ 0.9632 ใช้อัลกอริทึม PCA 0.9364 ใช้อัลกอริทึม Kernel PCA
0.9063 ใช้อัลกอริทึม AUTOENCODER 1 ใช้อัลกอริทึม tSNE 0.8160 ผลการทำนายของแบบจำลอง Decision Tree ได้ค่า accuracy ในกรณีไม่ใช้อัลกอริทึมสกัดคุณลักษณะ 0.9464 ใช้อัลกอริทึม
PCA 0.8929 ใช้อัลกอริทึม Kernel PCA 0.8695 ใช้อัลกอริทึม AUTOENCODER 1 ใช้อัลกอริทึม tSNE 0.8963 และในแบบจำลอง Random Forest ได้ค่า accuracy ในกรณีไม่ใช้อัลกอริทึมสกัดคุณลักษณะ 0.9331 ใช้อัลกอริทึม PCA 0.9230 ใช้อัลกอริทึม Kernel PCA 0.9163 ใช้อัลกอริทึม AUTOENCODER 1 ใช้อัลกอริทึม tSNE 0.8929 ตามลำดับ



ภาพประกอบที่ 22 แสดงผลการทำนายของชุดข้อมูล Credit Card Fraud Detection โดยใช้
อัลกอริทึมสกัดคุณลักษณะอื่นร่วมกับ tSNE

จากภาพประกอบที่ 11 แสดงผลการทำนายของชุดข้อมูล Credit Card Fraud Detection โดยใช้อัลกอริทึมสกัดคุณลักษณะอื่นร่วมกับ tSNE พบว่าในผลการทำนายของแบบจำลอง Logistic Regression ได้ค่า accuracy เมื่อใช้อัลกอริทึม PCA ร่วมกับ tSNE 0.9464 ใช้อัลกอริทึม Kernel PCA ร่วมกับ tSNE 0.8795 ใช้อัลกอริทึม AUTOENCODER ร่วมกับ tSNE 0.9985 ผลการทำนายของแบบจำลอง Support Vector Machine ได้ค่า accuracy เมื่อใช้อัลกอริทึม PCA ร่วมกับ tSNE 0.9364 ใช้อัลกอริทึม Kernel PCA ร่วมกับ tSNE 0.8862 ใช้อัลกอริทึม AUTOENCODER ร่วมกับ tSNE 1 ผลการทำนายของแบบจำลอง Decision Tree ได้ค่า accuracy เมื่อใช้อัลกอริทึม PCA ร่วมกับ tSNE 0.9030 ใช้อัลกอริทึม Kernel PCA ร่วมกับ tSNE 0.8896 ใช้อัลกอริทึม AUTOENCODER ร่วมกับ tSNE 1 และในแบบจำลอง Random Forest ได้ค่า accuracy เมื่อใช้อัลกอริทึม PCA ร่วมกับ tSNE 0.9431 ใช้อัลกอริทึม Kernel PCA ร่วมกับ tSNE 0.9264 ใช้อัลกอริทึม AUTOENCODER ร่วมกับ tSNE 1 ตามลำดับ

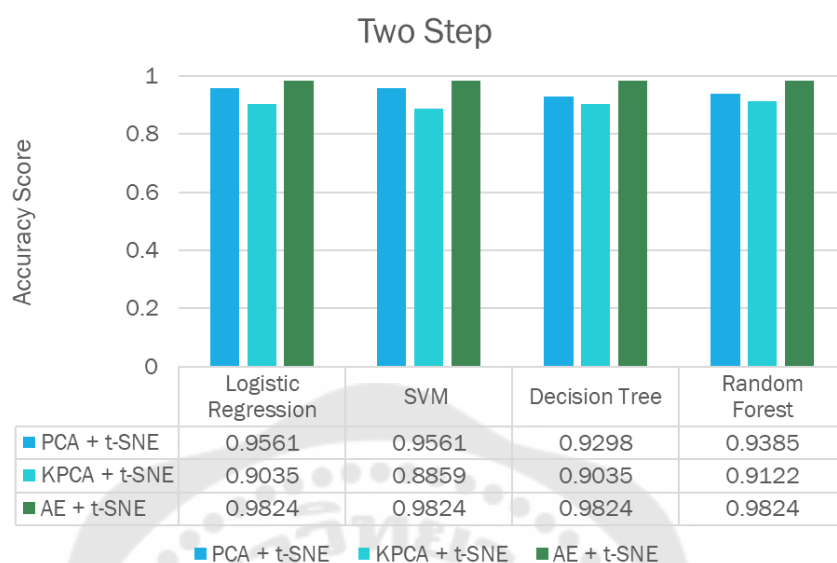
ผลลัพธ์การเปรียบเทียบประสิทธิภาพของแบบจำลองของชุดข้อมูล Breast Cancer

จากการทดลองเพื่อทำนายว่าเป็นข้อมูลการเป็นมะเร็งเต้านมหรือไม่ โดยใช้แบบจำลอง Logistic Regression, Support Vector Machine, Decision Tree และ Random Forest ได้ผลลัพธ์ดังนี้



ภาพประกอบที่ 23 แสดงผลการทำนายของชุดข้อมูล Breast Cancer โดยใช้อัลกอริทึมสกัดคุณลักษณะเดียว

จากภาพประกอบที่ 12 แสดงผลการทำนายของชุดข้อมูล Breast Cancer โดยใช้ อัลกอริทึมสกัดคุณลักษณะเดียว พบว่าในผลการทำนายของแบบจำลอง Logistic Regression ได้ ค่า accuracy ในกรณีไม่ใช้ อัลกอริทึมสกัดคุณลักษณะ 0.9824 ใช้ อัลกอริทึม PCA 0.9473 ใช้ อัลกอริทึม Kernel PCA 0.9122 ใช้ อัลกอริทึม AUTOENCODER 0.9736 ใช้ อัลกอริทึม tSNE 0.8947 ผลการทำนายของแบบจำลอง Support Vector Machine ได้ค่า accuracy ในกรณีไม่ใช้ อัลกอริทึมสกัดคุณลักษณะ 0.9736 ใช้ อัลกอริทึม PCA 0.9385 ใช้ อัลกอริทึม Kernel PCA 0.9210 ใช้ อัลกอริทึม AUTOENCODER 0.9561 ใช้ อัลกอริทึม tSNE 0.9035 ผลการทำนายของ แบบจำลอง Decision Tree ได้ค่า accuracy ในกรณีไม่ใช้ อัลกอริทึมสกัดคุณลักษณะ 0.9385 ใช้ อัลกอริทึม PCA 0.9298 ใช้ อัลกอริทึม Kernel PCA 0.9035 ใช้ อัลกอริทึม AUTOENCODER 0.9561 ใช้ อัลกอริทึม tSNE 0.9035 และในแบบจำลอง Random Forest ได้ค่า accuracy ใน กรณีไม่ใช้ อัลกอริทึมสกัดคุณลักษณะ 0.9385 ใช้ อัลกอริทึม PCA 0.9473 ใช้ อัลกอริทึม Kernel PCA 0.9385 ใช้ อัลกอริทึม AUTOENCODER 0.9824 ใช้ อัลกอริทึม tSNE 0.9035 ตามลำดับ



ภาพประกอบที่ 24 แสดงผลการทำนายของชุดข้อมูล Breast Cancer โดยใช้อัลกอริทึมสกัดคุณลักษณะอื่นร่วมกับ tSNE

จากภาพประกอบที่ 13 แสดงผลการทำนายของชุดข้อมูล Breast Cancer โดยใช้ อัลกอริทึมสกัดคุณลักษณะอื่นร่วมกับ tSNE พบว่าในผลการทำนายของแบบจำลอง Logistic Regression ได้ค่า accuracy เมื่อใช้อัลกอริทึม PCA ร่วมกับ tSNE 0.9561 ใช้อัลกอริทึม Kernel PCA ร่วมกับ tSNE 0.9035 ใช้อัลกอริทึม AUTOENCODER ร่วมกับ tSNE 0.9824 ผลการทำนายของแบบจำลอง Support Vector Machine ได้ค่า accuracy เมื่อใช้อัลกอริทึม PCA ร่วมกับ tSNE 0.9561 ใช้อัลกอริทึม Kernel PCA ร่วมกับ tSNE 0.8859 ใช้อัลกอริทึม AUTOENCODER ร่วมกับ tSNE 0.9824 ผลการทำนายของแบบจำลอง Decision Tree ได้ค่า accuracy เมื่อใช้อัลกอริทึม PCA ร่วมกับ tSNE 0.9298 ใช้อัลกอริทึม Kernel PCA ร่วมกับ tSNE 0.9035 ใช้อัลกอริทึม AUTOENCODER ร่วมกับ tSNE 0.9824 และในแบบจำลอง Random Forest ได้ค่า accuracy เมื่อใช้อัลกอริทึม PCA ร่วมกับ tSNE 0.9385 ใช้อัลกอริทึม Kernel PCA ร่วมกับ tSNE 0.9122 ใช้อัลกอริทึม AUTOENCODER ร่วมกับ tSNE 0.9824 ตามลำดับ

ผลลัพธ์การเปรียบเทียบประสิทธิภาพของแบบจำลองของชุดข้อมูล Heart Disease

จากการทดลองเพื่อทำนายว่าเป็นข้อมูลการเป็นโรคหัวใจหรือไม่ โดยใช้แบบจำลอง Logistic Regression, Support Vector Machine, Decision Tree และ Random Forest ได้ผลลัพธ์ดังนี้



ภาพประกอบที่ 25 แสดงผลการทำนายของชุดข้อมูล Heart Disease โดยใช้อัลกอริทึมสกัดคุณลักษณะเดียว

จากภาพประกอบที่ 14 แสดงผลการทำนายของชุดข้อมูล Heart Disease โดยใช้ อัลกอริทึมสกัดคุณลักษณะเดียว พบว่าในผลการทำนายของแบบจำลอง Logistic Regression ได้ ค่า accuracy ในกรณีไม่ใช้อัลกอริทึมสกัดคุณลักษณะ 0.7704 ใช้อัลกอริทึม PCA 0.8524 ใช้ อัลกอริทึม Kernel PCA 0.8360 ใช้อัลกอริทึม AUTOENCODER 0.8196 ใช้อัลกอริทึม tSNE 0.6885 ผลการทำนายของแบบจำลอง Support Vector Machine ได้ค่า accuracy ในกรณีไม่ใช้ อัลกอริทึมสกัดคุณลักษณะ 0.8196 ใช้อัลกอริทึม PCA 0.8524 ใช้อัลกอริทึม Kernel PCA 0.836 ใช้อัลกอริทึม AUTOENCODER 0.8196 ใช้อัลกอริทึม tSNE 0.7377 ผลการทำนายของ แบบจำลอง Decision Tree ได้ค่า accuracy ในกรณีไม่ใช้อัลกอริทึมสกัดคุณลักษณะ 0.8032 ใช้ อัลกอริทึม PCA 0.6885 ใช้อัลกอริทึม Kernel PCA 0.6557 ใช้อัลกอริทึม AUTOENCODER 0.7540 ใช้อัลกอริทึม tSNE 0.5245 และในแบบจำลอง Random Forest ได้ค่า accuracy ใน กรณีไม่ใช้อัลกอริทึมสกัดคุณลักษณะ 0.7704 ใช้อัลกอริทึม PCA 0.7540 ใช้อัลกอริทึม Kernel PCA 0.8196 ใช้อัลกอริทึม AUTOENCODER 0.8852 ใช้อัลกอริทึม tSNE 0.5573 ตามลำดับ



ภาพประกอบที่ 26 แสดงผลการทำนายของชุดข้อมูล Heart Disease โดยใช้อัลกอริทึมสกัดคุณลักษณะอื่นร่วมกับ tSNE

จากภาพประกอบที่ 15 แสดงผลการทำนายของชุดข้อมูล Heart Disease โดยใช้ อัลกอริทึมสกัดคุณลักษณะอื่นร่วมกับ tSNE พบว่าในผลการทำนายของแบบจำลอง Logistic Regression ได้ค่า accuracy เมื่อใช้อัลกอริทึม PCA ร่วมกับ tSNE 0.8524 ใช้อัลกอริทึม Kernel PCA ร่วมกับ tSNE 0.8360 ใช้อัลกอริทึม AUTOENCODER ร่วมกับ tSNE 0.7704 ผลการทำนายของแบบจำลอง Support Vector Machine ได้ค่า accuracy เมื่อใช้อัลกอริทึม PCA ร่วมกับ tSNE 0.8688 ใช้อัลกอริทึม Kernel PCA ร่วมกับ tSNE 0.8524 ใช้อัลกอริทึม AUTOENCODER ร่วมกับ tSNE 0.7213 ผลการทำนายของแบบจำลอง Decision Tree ได้ค่า accuracy เมื่อใช้อัลกอริทึม PCA ร่วมกับ tSNE 0.7377 ใช้อัลกอริทึม Kernel PCA ร่วมกับ tSNE 0.6885 ใช้อัลกอริทึม AUTOENCODER ร่วมกับ tSNE 0.8524 และในแบบจำลอง Random Forest ได้ค่า accuracy เมื่อใช้อัลกอริทึม PCA ร่วมกับ tSNE 0.7868 ใช้อัลกอริทึม Kernel PCA ร่วมกับ tSNE 0.7540 ใช้อัลกอริทึม AUTOENCODER ร่วมกับ tSNE 0.9016 ตามลำดับ

บทที่ 5

สรุป อภิปรายผล และข้อเสนอแนะ

ในการวิจัยการศึกษาผลของการใช้อัลกอริทึมสกัดคุณลักษณะต่ออัลกอริทึม tSNE ผู้วิจัยได้วัดประสิทธิภาพของแบบจำลองในการใช้อัลกอริทึมสกัดคุณลักษณะในแต่ละอัลกอริทึมเพื่อนำมาเปรียบเทียบและสรุปผลโดยสามารถแบ่งหัวข้อในการสรุปผลได้ดังต่อไปนี้

สรุปผลการวิจัย

อภิปรายผล

ข้อเสนอแนะ

สรุปผลการวิจัย

ในปัจจุบันอัลกอริทึมสกัดคุณลักษณะ tSNE ได้รับความนิยมอย่างมากในการใช้ลดจำนวนมิติของชุดข้อมูลและการแสดงภาพของข้อมูล ทว่าการใช้อัลกอริทึม tSNE กับข้อมูลเพื่อลดจำนวนมิติลงก่อนทำการจำแนกข้อมูลจะทำให้ประสิทธิภาพการจำแนกข้อมูลลดลงอย่างมากเนื่องจากอัลกอริทึม tSNE ไม่สามารถรักษาคุณสมบัติโดยรวมของข้อมูลไว้ได้ ซึ่งหากปัญหาดังกล่าวนั้นสามารถแก้ไขได้ด้วยการเพิ่มประสิทธิภาพให้กับการใช้อัลกอริทึม tSNE เพื่อลดมิติได้มากขึ้นจะช่วยให้การทำงานด้วยเทคนิคการเรียนรู้ของเครื่องมีประสิทธิภาพยิ่งขึ้น

ในการวิจัยนี้ผู้วิจัยเลือกศึกษาประสิทธิภาพของการจำแนกประเภทแบบสองกลุ่มโดยใช้อัลกอริทึมสกัดคุณลักษณะ PCA, Kernel PCA, AUTOENCODER ทำงานร่วมกับอัลกอริทึม tSNE และใช้ชุดข้อมูลทั้งหมด 3 ชุด และนำมาสร้างแบบจำลอง Logistic Regression, Support Vector Machine, Decision Tree และ Random Forest เพื่อเป็นแนวทางในการเพิ่มประสิทธิภาพให้กับการใช้อัลกอริทึมสกัดคุณลักษณะ tSNE โดยสามารถสรุปการวัดประสิทธิภาพของแต่ละแบบจำลองได้ดังนี้

ตารางที่ 1 การเปรียบเทียบประสิทธิภาพอัลกอริทึมสกัดคุณลักษณะของชุดข้อมูล Credit Card Fraud Detection

อัลกอริทึม	Logistic Regression	Support Vector Machine	Decision Tree	Random Forest
Baseline	0.96	0.96	0.94	0.93
PCA	0.93	0.93	0.89	0.92
KPCA	0.88	0.91	0.91	0.94
AUTOENCODER	1	1	1	1
tSNE	0.79	0.81	0.89	0.89
PCA และ tSNE	0.94	0.93	0.90	0.94
KPCA และ tSNE	0.82	0.82	0.91	0.94
AE และ tSNE	0.99	1	1	1

จากผลการทดลองสรุปได้ว่าในชุดข้อมูล Credit Card Fraud Detection การใช้
อัลกอริทึม AUTOENCODER ร่วมกับ tSNE ให้ประสิทธิภาพดีที่สุดในทุกแบบจำลอง โดยค่าที่ได้
จากแบบจำลอง Logistic Regression คือ 0.99 Support Vector Machine 1 Decision Tree 1
และ Random Forest 1

ตารางที่ 2 การเปรียบเทียบประสิทธิภาพอัลกอริทึมสกัดคุณลักษณะของชุดข้อมูล Breast Cancer Wisconsin

อัลกอริทึม	Logistic Regression	Support Vector Machine	Decision Tree	Random Forest
Baseline	0.98	0.97	0.93	0.93
PCA	0.94	0.93	0.92	0.94
KPCA	0.91	0.92	0.90	0.93
AUTOENCODER	0.97	0.95	0.95	0.98
tSNE	0.89	0.90	0.90	0.90
PCA และ tSNE	0.95	0.95	0.92	0.93
KPCA และ tSNE	0.90	0.88	0.90	0.93
AE และ tSNE	0.98	0.98	0.98	0.98

จากผลการทดลองสรุปได้ว่าในชุดข้อมูล Breast Cancer Wisconsin การใช้อัลกอริทึม AUTOENCODER ร่วมกับ tSNE ให้ประสิทธิภาพดีที่สุดในทุกแบบจำลอง โดยค่าที่ได้จากแบบจำลอง Logistic Regression คือ 0.98 Support Vector Machine 0.98 Decision Tree 0.98 และ Random Forest 0.98

ตารางที่ 3 การเปรียบเทียบประสิทธิภาพอัลกอริทึมสังกัดคุณลักษณะของชุดข้อมูล Heart Disease

อัลกอริทึม	Logistic Regression	Support Vector Machine	Decision Tree	Random Forest
Baseline	0.77	0.81	0.80	0.77
PCA	0.85	0.85	0.68	0.75
KPCA	0.83	0.83	0.65	0.81
AUTOENCODER	0.81	0.81	0.75	0.88
tSNE	0.68	0.73	0.52	0.55
PCA และ tSNE	0.85	0.86	0.73	0.78
KPCA และ tSNE	0.83	0.85	0.68	0.75
AE และ tSNE	0.77	0.72	0.85	0.90

จากผลการทดลองสรุปได้ว่าในชุดข้อมูล Heart Disease การใช้อัลกอริทึม PCA ร่วมกับ tSNE ให้ประสิทธิภาพดีที่สุดในแบบจำลอง Logistic Regression และ Random Forest โดยค่าที่ได้จากแบบจำลอง Logistic Regression คือ 0.85 Support Vector Machine 0.86 ส่วนการใช้ อัลกอริทึม AUTOENCODER ร่วมกับ tSNE ให้ประสิทธิภาพดีที่สุดในแบบจำลอง Decision Tree และ Random Forest โดยค่าที่ได้จากแบบจำลอง Decision Tree 0.85 และ Random Forest 0.90

อภิปรายผล

ในการศึกษาประสิทธิภาพของแบบจำลองพบว่าการใช้อัลกอริทึมสังกัดคุณลักษณะอื่น ทำงานร่วมกับอัลกอริทึม tSNE แบบสองชั้นสามารถช่วยให้ประสิทธิภาพของแบบจำลองดีขึ้น อย่างมีนัยสำคัญมากกว่าการใช้อัลกอริทึม tSNE เพียงอัลกอริทึมเดียวในการลดมิติข้อมูลจากการ ทดลองในทุกชุดข้อมูล

ตารางที่ 4 แสดงประสิทธิภาพของแบบจำลองจำแนกประเภทเมื่อใช้อัลกอริทึมคุณลักษณะแบบ สองชั้นเทียบกับการใช้อัลกอริทึมเดียว

ชุดข้อมูล/อัลกอริทึม	Credit Card Fraud Detection	Breast Cancer Wisconsin	Heart Disease
tSNE / PCA + tSNE	0.79 / 0.94 (+18.9%)	0.89 / 0.95 (+6.7%)	0.68 / 0.85 (+25%)
tSNE / KPCA + tSNE	0.94 / 0.94 (0%)	0.93 / 0.93 (0%)	0.83 / 0.85 (+2.4%)
tSNE / AE + tSNE	1 / 1 (0%)	0.95 / 0.98 (+3.1%)	0.75 / 0.85 (+13%)

จากตารางที่ 4 จะเห็นได้ว่าการใช้อัลกอริทึมสกัดคุณลักษณะแบบสองชั้นให้ผลลัพธ์ที่ดีกว่าการใช้อัลกอริทึม tSNE เพียงอัลกอริทึมเดียว โดยการใช้อัลกอริทึม PCA สามารถช่วยเพิ่มประสิทธิภาพให้กับอัลกอริทึม tSNE ได้ในทุกชุดข้อมูล การใช้อัลกอริทึม KPCA สามารถช่วยเพิ่มประสิทธิภาพให้กับอัลกอริทึม tSNE ได้ในชุดข้อมูล Heart Disease การใช้อัลกอริทึม AUTOENCODER สามารถช่วยเพิ่มประสิทธิภาพให้กับอัลกอริทึม tSNE ได้ในชุดข้อมูล Breast Cancer Wisconsin และ Heart Disease

ประสิทธิภาพในการใช้อัลกอริทึม AUTOENCODER เพียงอัลกอริทึมเดียวเทียบกับการใช้อัลกอริทึม AUTOENCODER ร่วมกับ tSNE การทดลองในชุดข้อมูล Credit Card Fraud Detection พบว่าประสิทธิภาพใกล้เคียงกัน จากข้อมูลที่เป็นผลลัพธ์จากอัลกอริทึม AUTOENCODER มีจำนวน 50 มิติ มากกว่าการใช้ AUTOENCODER ร่วมกับ tSNE ที่มีจำนวน 2 มิติ ทั้งนี้จะเห็นได้ว่า แม้การทดลองด้วยการใช้อัลกอริทึม AUTOENCODER เพียงอย่างเดียว จะมีการใช้มิติมากกว่าเป็นอย่างมาก แต่มิได้มีประสิทธิภาพดีกว่าการใช้ AUTOENCODER ร่วมกับ อัลกอริทึม tSNE มากนัก

ในทางกลับกันการใช้อัลกอริทึม AUTOENCODER ร่วมกับ tSNE ให้ประสิทธิภาพที่ดีกว่าหมายความว่าอัลกอริทึม tSNE ช่วยให้ข้อมูลเกาะกลุ่มกันได้ดีขึ้นแม้จะเหลือมิติน้อยกว่า ดังผลลัพธ์ที่เห็นในชุดข้อมูล Breast Cancer พบว่าประสิทธิภาพของการใช้อัลกอริทึม AUTOENCODER อัลกอริทึมเดียวเทียบกับการใช้อัลกอริทึม AUTOENCODER ร่วมกับ tSNE ให้ผลลัพธ์ที่ใกล้เคียงกันดังนั้นอาจเลือกใช้เพียงอัลกอริทึม AUTOENCODER เพียงอย่างเดียวแต่หากความมุ่งหมายของผู้วิจัยต้องการศึกษาอัลกอริทึมที่ช่วยเพิ่มประสิทธิภาพให้อัลกอริทึม tSNE จึงให้ผลลัพธ์ของการใช้อัลกอริทึมร่วมกันเป็นหลัก

จากผลการทดลองนี้แสดงให้เห็นว่าการใช้ AUTOENCODER เพื่อสกัดคุณลักษณะของข้อมูลออกมาชั้นหนึ่งก่อนมีผลทำให้ข้อมูลมีการแบ่งกลุ่มออกได้อย่างมีประสิทธิภาพมากขึ้น ทั้งนี้ อาจเป็นผลเนื่องจาก AUTOENCODER ช่วยทำให้ข้อมูลที่มีลักษณะใกล้เคียงกัน จับกลุ่มกันได้ดีขึ้น ทำให้ผลลัพธ์การจำแนกประเภทเกือบทุกแบบจำลองในทุกชุดข้อมูลดีขึ้นทั้งสิ้นเนื่องจาก AUTOENCODER นั้นมีลักษณะที่บีบเล็กลงตรงกลางที่เรียกว่า Code หรือ Latent Layer ซึ่งเป็น

ส่วนที่เก็บแก่นใจความสำคัญของข้อมูลที่ถูกสกัดมาจากชั้น Encoder และชั้น Decoder จะทำการขยายใจความสำคัญของข้อมูลจาก Code ให้มีความสมบูรณ์มากที่สุด อีกสิ่งหนึ่งคือ AUTOENCODER เป็น Neural Network ที่ใช้คุณสมบัติ Backpropagation เพื่อปรับค่า Loss Function เพื่อพยายามสร้างข้อมูล Output ให้เหมือนกับข้อมูล Input มากที่สุด รวมไปถึง AUTOENCODER มี Activation Function เป็นแบบไม่เชิงเส้นจึงสามารถทำงานกับข้อมูลที่มีโครงสร้างแบบไม่เชิงเส้นได้

เปรียบเทียบประสิทธิภาพการใช้อัลกอริทึมสกัดคุณลักษณะต่าง ๆ ร่วมกับอัลกอริทึม tSNE ในการศึกษาได้ใช้อัลกอริทึมสกัดคุณลักษณะ PCA, KPCA และ AUTOENCODER ร่วมกับ tSNE เพื่อเปรียบเทียบประสิทธิภาพ จากผลการทดลอง พบว่า โดยรวมแล้ว AUTOENCODER ให้ผลลัพธ์ที่ดีที่สุดในเกือบทุกชุดข้อมูล ที่ใช้ตัวจำแนกที่แตกต่างกัน ดังที่แสดงในตารางที่ 1, 2 และ 3

ทั้งนี้สาเหตุที่อัลกอริทึม AUTOENCODER ดีที่สุดเมื่อเปรียบเทียบกับอัลกอริทึม PCA และ KPCA อาจเนื่องจากอัลกอริทึม PCA เป็นอัลกอริทึมแบบเชิงเส้นจึงไม่สามารถจัดการแยกกลุ่มกับข้อมูลที่มีความซับซ้อนมากได้ดีนัก ส่วนอัลกอริทึม KPCA แม้ว่าจะสามารถทำงานกับข้อมูลประเภทไม่เชิงเส้นได้ แต่อัลกอริทึม KPCA เป็นสกัดคุณลักษณะโดยปราศจากการปรับเข้ากับข้อมูลทำให้ได้ประสิทธิภาพไม่ดีนัก ในทางตรงข้ามอัลกอริทึม AUTOENCODER จะมีการปรับค่าความผิดพลาด (Loss Function) จากเรียนรู้ข้อมูลจากที่ผิดพลาด ทำให้สามารถทำงานข้อมูลที่มีความซับซ้อนได้ดีที่สุด ทำให้มีประสิทธิภาพในการทำงานร่วมกับอัลกอริทึม tSNE สูงสุด

ทั้งนี้ในการศึกษาในขั้นต่อไปอาจขยายขอบเขตการศึกษา โดยเพิ่มชุดข้อมูลที่มีความหลากหลายจากงานที่แตกต่างกันในสาขาต่าง ๆ เพื่อศึกษาความเหมาะสมของการใช้อัลกอริทึม AUTOENCODER ในข้อมูลด้านต่าง ๆ เช่น ข้อมูลทางการแพทย์ ข้อมูลด้านภาพถ่ายดาวเทียม หรือข้อมูลพฤติกรรมการทำงานในระบบคอมพิวเตอร์ เป็นต้น

ข้อเสนอแนะ

1. หากมีการทำการสำรวจข้อมูลหรือ Exploration Data Analysis อย่างละเอียดจะช่วยเพิ่ม Information ของข้อมูลมากขึ้น
2. หากมีการปรับแต่ง Parameter ของอัลกอริทึมสกัดคุณลักษณะและแบบจำลองอย่างละเอียดอาจทำให้ประสิทธิภาพการสกัดคุณลักษณะข้อมูลแบบสองชั้นนอกเหนือจากอัลกอริทึม AUTOENCODER ดีขึ้นกว่านี้

บรรณานุกรม

- Altman, N., & Krzywinski, M. (2018). The curse (s) of dimensionality. *Nat Methods*, 15, 399-400.
- Andras Janosi, W. S., Matthias Pfisterer, Robert Detrano. (1988). *Heart Disease Data Set*. Bruxelles), M. L. G. o. U. U. L. d. (2018). *Credit Card Fraud Detection*. Retrieved from: <https://www.kaggle.com/mlg-ulb/creditcardfraud>
- Chandrayan, P. (2015). Logistic Regression For Dummies: A Detailed Explanation. Retrieved from <https://towardsdatascience.com/logistic-regression-for-dummies-a-detailedexplanation-9597f76edf46>
- Dr. William H. Wolberg, W. N. S., Olvi L. Mangasarian. (1995). *Breast Cancer Wisconsin (Diagnostic) Data Set*.
- Heaton, J. (2016). *An empirical analysis of feature engineering for predictive modeling*. Paper presented at the SoutheastCon 2016.
- Jamal, A., Handayani, A., Septiandri, A. A., Ripmiation, E., & Effendi, Y. (2018). Dimensionality Reduction using PCA and K-Means Clustering for Breast Cancer Prediction. *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, 192.
- Kobak, D., & Berens, P. (2019). The art of using tSNE for single-cell transcriptomics. *Nature Communications*, 10(1), 5416.
- Kurita, K. (2018). Paper Dissected: "Visualizing Data using tSNE" Explained. Retrieved from <https://mlexplained.com/2018/09/14/paper-dissected-visualizing-data-using-tSNE-explained/>
- Navlani, A. (2019). Support Vector Machines with Scikit-learn. Retrieved from <https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python#svm>
- Novakovic, J. (2016). Toward optimal feature selection using ranking methods and classification algorithms. *Yugoslav Journal of Operations Research*, 21(1).
- Song, W., Wang, L., Liu, P., & Choo, K.-K. R. (2018). Improved tSNE based manifold dimensional reduction for remote sensing data processing. *Multimedia Tools and*

Applications, 78(4), 4311-4326.

van der Maaten, L., & Hinton, G. (2008). Visualizing Data using tSNE. *Journal of Machine Learning Research*, 9, 2579-2605.

Wang, Y., Yao, H., & Zhao, S. (2016). Auto-encoder based dimensionality reduction. *Neurocomputing*, 184, 232-242.

Zheng, A., & Casari, A. (2018). *Feature engineering for machine learning: principles and techniques for data scientists*: " O'Reilly Media, Inc."



ประวัติผู้เขียน

ชื่อ-สกุล	นายนรเศรษฐ์ กรุดสมัย
วัน เดือน ปี เกิด	29 พฤษภาคม 2532
สถานที่เกิด	กรุงเทพฯ
วุฒิการศึกษา	พ.ศ. 2550 ศิลปบัณฑิต สาขาดนตรีสากล จาก มหาวิทยาลัยเทคโนโลยีราชมงคลธัญบุรี พ.ศ. 2561 วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ จาก มหาวิทยาลัยศรีนครินทรวิโรฒ

