การจำแนกกลุ่มผู้เป็นโรคหัวใจ
ด้วยแบบจำลองแยกประเภทของข้อมูล (Classification Model)

โดย

6504610152 โชติวิทย์ เกษมชัยนันท์

6504680247 ธัญเทพ เชี่ยวชาญช่าง

เสนอ

ผู้ช่วยศาสตราจารย์ ดร.ธีรวุฒิ ศรีพินิจ

งานศึกษานี้เป็นส่วนหนึ่งของวิชา ศ.200 วิทยาศาสตร์ข้อมูลสำหรับการวิเคราะห์เศรษฐกิจ

ภาคเรียนที่ 2 ปีการศึกษา 2566

คณะเศรษฐศาสตร์ มหาวิทยาลัยธรรมศาสตร์

บทคัดย่อ

งานศึกษาในวิชาวิทยาศาสตร์ข้อมูลสำหรับการวิเคราะห์เศรษฐกิจ (ศ. 200) ฉบับนี้ มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองแบ่งแยกประเภท (Classification Model) ที่ดีที่สุด และระบุน้ำหนักของความดันโลหิต อัตราการเต้นของหัวใจ และปัจจัยอื่น ๆ ที่ใช้ในการคาดการณ์ว่าเป็นหรือไม่เป็นผู้ป่วยโรคหัวใจ โดยข้อมูลของกลุ่มตัวอย่างที่ใช้เป็นข้อมูลในการสร้างแบบจำลอง ได้จากการหาข้อมูลโดยการสืบค้นบนอินเทอร์เน็ต คือ ชุดข้อมูล “Heart Disease Dataset” ซึ่งประกอบด้วยข้อมูลจำนวน 1190 ตัวอย่าง การสร้างแบบจำลองในการวิจัยนี้ใช้เทคนิคการจำแนกประเภทข้อมูลจำนวน 2 วิธี ได้แก่ 1. แบบจำลองต้นไม้ตัดสินใจ (Decision Tree Model) และแบบจำลองป่าสุ่ม (Random Forest Model) ด้วยโปรแกรม Python ประเมินประสิทธิภาพของแบบจำลองด้วยเกณฑ์ Accuracy, Score, Precision Score, Recall Score และ F1-Score โดยอาศัยข้อมูลจาก Confusion Matrix ผลการเปรียบเทียบประสิทธิภาพของแบบจำลอง พบว่า แบบจำลองที่ดีที่สุด หรือมีประสิทธิภาพสูงสุด คือแบบจำลองป่าไม้สุ่ม (Random Forest Model) โดยใช้พารามิเตอร์ criterion= 'gini', max_depth = 9, n_estimators = 7 และมีตัวแปรอิสระ 10 ตัว ได้แก่ อายุ เพศ ลักษณะการเจ็บหน้าอก ความดันโลหิตขณะพัก ระดับของคอเลสเตอรอลในเลือด คลื่นไฟฟ้าหัวใจขณะพัก อัตราการเต้นของหัวใจสูงสุด การมีอาการเจ็บหน้าอกขณะออกกำลังกาย ระดับการยกตัวของ ST segment เทียบกับ Isoelectric Line และลักษณะความชันของ ST segment โดยมีค่า Accuracy Score เท่ากับ 0.9248 ค่า Precision Score สำหรับกลุ่มที่ไม่เป็นโรคหัวใจเท่ากับ 0.9028 และเป็นโรคหัวใจเท่ากับ 0.9444 ค่า Recall Socre สำหรับกลุ่มที่ไม่เป็นโรคหัวใจเท่ากับ 0.9353 และเป็นโรคหัวใจเท่ากับ 0.9162 ค่า F1-score สำหรับกลุ่มที่ไม่เป็นโรคหัวใจเท่ากับ 0.9187 และเป็นโรคหัวใจเท่ากับ 0.9301

สารบัญ

หน้า

บทคัดย่อ ก

สารบัญ ข

สารบัญรูปภาพ จ

สารบัญตาราง ช

บทที่ 1 บทนำ

1.1 ที่มาและความสำคัญ 1

1.2 วัตถุประสงค์ 2

1.3 ขอบเขตงานศึกษา 2

บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 Machine Learning from SKLearn 3

2.1.1 การ Train และ Test ใน Machine Learning from SKLearn ด้วยภาษา Python 3

2.1.2 วิธีการ Train ใน Machine Learning 4

2.1.3 วิธีการ Test ใน Machine Learning 4

2.2 Classification Model 5

2.3 Regression 7

2.4 Linear Regression Model 8

2.5 Decision Tree Model 8

2.5.1 ประเภทของ Decision Tree 9

2.5.1.1 Regression Tree 9

2.5.1.2 Classification Tree 9

2.5.2 วิธีการใช้งาน Decision Tree ใน Python 10

2.6 Random Forest Model 11

2.7 Confusion Matrix 12

2.7.1 Accuracy 13

2.7.2 Precision 13

2.7.3 Recall 14
2.7.4 F1-Score 15

บทที่ 3 วิธีการดำเนินการวิจัย
3.1 การรวบรวมข้อมูลเพื่อใช้สร้างแบบจำลอง 15
3.2 การกำหนดตัวแปร 15
3.2.1 การกำหนดตัวแปรอิสระ 15
3.2.2 การกำหนดตัวแปรตาม 15
3.3 การเตรียมข้อมูล 15
3.3.1 กระบวนการตรวจสอบและจัดการข้อมูลที่หายไป 15
3.3.2 กระบวนการแปลงข้อมูลเพื่อให้สามารถนำไปสร้างแบบจำลองได้ 16
3.4 การวิเคราะห์ข้อมูล 16
3.4.1 ตัวแปรเชิงปริมาณ (Continuous Variable) 16
3.4.2 ตัวแปรเชิงคุณภาพ (Ordinal หรือ Nominal Scale Variable) 16
3.4.3 เปรียบเทียบลักษณะของข้อมูล เมื่อค่าของตัวแปรเป้าหมาย (Target Class) ต่างกัน16
3.5 การสร้างแบบจำลองต้นไม้ตัดสินใจ (Decision Tree Model) พร้อมประเมินประสิทธิภาพ 16
3.5.1 การหาน้ำหนักความสำคัญของตัวแปร (Feature Importance) 16
3.5.2 การหาชุดของตัวแปรและพารามิเตอร์ของแบบจำลองที่ดีที่สุด 17
3.5.3 ประเมินประสิทธิภาพของแบบจำลอง 17
3.6 การสร้างแบบจำลองป่าไม้สุ่ม (Random Forest Model) พร้อมประเมินประสิทธิภาพ 17
3.6.1 การหาน้ำหนักความสำคัญของตัวแปร (Feature Importance) 17
3.6.2 การหาชุดของตัวแปรและพารามิเตอร์ของแบบจำลองที่ดีที่สุด 18
3.6.3 ประเมินประสิทธิภาพของแบบจำลอง 18
3.7 สรุป และอภิปรายผล 19

บทที่ 4 ผลการดำเนินการวิจัย
4.1 ข้อมูลที่ได้จากการรวบรวมเพื่อใช้สร้างแบบจำลอง 20
4.2 ตัวแปรที่ได้จากการกำหนด 20
4.3 ผลการเตรียมข้อมูล 21
4.4 ผลการวิเคราะห์ข้อมูล 22

4.4.1 ตัวแปรเชิงปริมาณ (Continuous Variable) 22
4.4.2 ตัวแปรเชิงคุณภาพ (Ordinal หรือ Nominal Scale Variable) 25
4.4.3 เปรียบเทียบลักษณะของข้อมูล เมื่อค่าของตัวแปรเป้าหมาย (Target Class) ต่างกัน 29
4.5 แบบจำลองต้นไม้ตัดสินใจ (Decision Tree Model) และประสิทธิภาพของแบบจำลอง 31
4.5.1 น้ำหนักความสำคัญของตัวแปร (Feature Importance) 31
4.5.2 ชุดของตัวแปรและพารามิเตอร์ของแบบจำลองที่ดีที่สุด 32
4.5.2.1 ชุดของตัวแปร 32
4.5.2.2 แบบจำลองที่ดีที่สุดจากแต่ละชุดของตัวแปร 33
4.5.3 ประสิทธิภาพของแบบจำลอง 33
4.6 แบบจำลองป่าไม้สุ่ม (Random Forest Model) และประสิทธิภาพของแบบจำลอง 35
4.6.1 น้ำหนักความสำคัญของตัวแปร (Feature Importance) 35
4.6.2 ชุดของตัวแปรและพารามิเตอร์ของแบบจำลองที่ดีที่สุด 36
4.6.2.1 ชุดของตัวแปร 36
4.6.2.2 แบบจำลองที่ดีที่สุดจากแต่ละชุดของตัวแปร 37
4.6.3 ประสิทธิภาพของแบบจำลอง 38
4.7 สรุป และอภิปรายผล 40

บทที่ 5 สรุปผลการวิจัย และข้อเสนอแนะ
5.1 สรุปผลการวิจัย 41
5.2 ข้อเสนอแนะ 42
บรรณานุกรม 43

สารบัญรูปภาพ

หน้า

รูปที่ 2.1 หลักการทำงานของ Classification Model (Datacamp, 2565) 5

รูปที่ 2.2 Binary Classification (Datacamp, 2565) 6

รูปที่ 2.3 Multi-Label Classification (Datacamp, 2565) 6

รูปที่ 2.4 Imbalanced Classification (Datacamp, 2565) 7

รูปที่ 2.5 Decision Tree Model (Teerawut Sripinit, 2567) 8

รูปที่ 2.6 แบบจำลอง Random Forest (IBM, ม.ป.ป.) 12

รูปที่ 2.7 Confusion Matrix (Medium, 2019) 13

รูปที่ 4.1 ตัวอย่างข้อมูลที่ได้จากการสืบค้นบนอินเทอร์เน็ตและนำไปสร้างแบบจำลอง 20

รูปที่ 4.2 ลักษณะการกระจายของตัวแปร age (อายุ) 23

รูปที่ 4.3 ลักษณะการกระจายของตัวแปร resting bp s (ความดันโลหิตขณะพัก) 23

รูปที่ 4.4 ลักษณะการกระจายของตัวแปร cholesterol (ระดับของคอเลสเตอรอลในเลือด) 24

รูปที่ 4.5 ลักษณะการกระจายของตัวแปร max heart rate (อัตราการเต้นของหัวใจสูงสุด) 24

รูปที่ 4.6 ลักษณะการกระจายของตัวแปร oldpeak (ระดับการยกตัวของ ST segment เทียบกับ Isoelectric Line) 25

รูปที่ 4.7 แผนภูมิวงกลมแสดงสัดส่วนความถี่ของตัวแปร sex 25

รูปที่ 4.8 แผนภูมิวงกลมแสดงสัดส่วนความถี่ของตัวแปร chest pain type 26

รูปที่ 4.9 แผนภูมิวงกลมแสดงสัดส่วนความถี่ของตัวแปร fasting blood sugar 27

รูปที่ 4.10 แผนภูมิวงกลมแสดงสัดส่วนความถี่ของตัวแปร resting ecg 27

รูปที่ 4.11 แผนภูมิวงกลมแสดงสัดส่วนความถี่ของตัวแปร exercise angina 28

รูปที่ 4.12 แผนภูมิวงกลมแสดงสัดส่วนความถี่ของตัวแปร ST slope 28

รูปที่ 4.13 แผนภูมิวงกลมแสดงสัดส่วนความถี่ของตัวแปร target 29

รูปที่ 4.14 Box plot ของตัวแปรเชิงปริมาณ เมื่อ target เป็น 0 และ 1 29

สารบัญตาราง

หน้า

ตารางที่ 4.1 ตัวแปรคุณลักษณะ ตัวแปรเป้าหมาย และเงื่อนไขที่ใช้ในการสร้างต้นไม้ตัดสินใจ 20

ตารางที่ 4.2 Data Type และ Missing Value ของข้อมูลก่อนและหลังการจัดการ 21

ตารางที่ 4.3 สถิติเชิงพรรณนา (Descriptive Statistics) ของตัวแปรเชิงปริมาณ 22

ตารางที่ 4.4 ตารางแจกแจงความถี่ของตัวแปร sex 25

ตารางที่ 4.5 ตารางแจกแจงความถี่ของตัวแปร chest pain type 26

ตารางที่ 4.6 ตารางแจกแจงความถี่ของตัวแปร fasting blood sugar 27

ตารางที่ 4.7 ตารางแจกแจงความถี่ของตัวแปร resting ecg 27

ตารางที่ 4.8 ตารางแจกแจงความถี่ของตัวแปร exercise angina 28

ตารางที่ 4.9 ตารางแจกแจงความถี่ของตัวแปร ST slope 28

ตารางที่ 4.10 ตารางแจกแจงความถี่ของตัวแปร target 29

ตารางที่ 4.11 ค่าเฉลี่ยของตัวแปรเชิงปริมาณ เมื่อ target เป็น 0 และ 1 29

ตารางที่ 4.12 ฐานนิยมของตัวแปรเชิงคุณภาพ เมื่อ Target เป็น 0 และ 1 30

ตารางที่ 4.13 น้ำหนักความสำคัญของตัวแปรคุณลักษณะแต่ละตัว (Feature Importance) เรียงจากมากไปน้อย 31

ตารางที่ 4.14 แบบจำลองที่ดีที่สุดจากแต่ละชุดของตัวแปร 33

ตารางที่ 4.15 ประสิทธิภาพของแบบจำลอง ทดสอบด้วย Training Data 33

ตารางที่ 4.16 ประสิทธิภาพของแบบจำลอง ทดสอบด้วย Testing Data 34

ตารางที่ 4.17 น้ำหนักความสำคัญของตัวแปรคุณลักษณะแต่ละตัว (Feature Importance) เรียงจากมากไปน้อย 36

ตารางที่ 4.18 แบบจำลองที่ดีที่สุดในแต่ละชุดของตัวแปรใน Random Forest 37

ตารางที่ 4.19 ประสิทธิภาพของแบบจำลอง ทดสอบด้วย Training Data 38

ตารางที่ 4.20 ประสิทธิภาพของแบบจำลอง ทดสอบด้วย Testing Data 39

ตารางที่ 5.1 การเปรียบเทียบประสิทธิภาพของแบบจำลองป่าไม้สุ่ม และแบบจำลองต้นไม้ตัดสินใจ 42

บทที่ 1

บทนำ

1.1 ที่มาและความสำคัญ

โรคหัวใจ (Heart Disease) เป็นกลุ่มคำที่มีความหมายครอบคลุมหลายโรคและภาวะที่ก่อให้เกิดความผิดปกติในการทำงานของหัวใจ เช่น โรคหลอดเลือดหัวใจ โรคลิ้นหัวใจตีบ โรคกล้ามเนื้อหัวใจ โรคเยื่อหุ้มหัวใจอักเสบ การติดเชื้อที่หัวใจ ภาวะหัวใจเต้นผิดปกติ ภาวะบกพร่องของหัวใจแต่กำเนิด ซึ่งผู้ป่วยโรคหัวใจอาจเกิดภาวะแทรกซ้อน อาทิ ภาวะหัวใจล้มเหลว โรคหลอดเลือดหัวใจ โดยหากมีอาการเจ็บหน้าอกร่วมกับหายใจถี่เหนื่อยง่าย หรือเป็นลมควรรีบพบแพทย์ให้เร็วที่สุด โดยเฉพาะอย่างยิ่งในกรณีที่มีสมาชิกในครอบครัวมีประวัติเป็นโรคหัวใจ อย่างไรก็ตามโรคหัวใจจะรักษาได้ผลดี หากตรวจพบโรคหัวใจตั้งแต่ระยะแรก ๆ (ไพศาล บุญศิริคำชัย, 2564)

ตั้งแต่ปีพุทธศักราช 2553 ถึง 2561 ในกลุ่มคนอายุ 30-69 ปี ประเทศไทยมีอัตราการเสียชีวิตจากโรคหัวใจและหลอดเลือดเป็นอันดับสองในหมู่โรคไม่ติดต่อเรื้อรัง รองจากโรคมะเร็ง โดยในปี 2561 มีอัตราการเสียชีวิตจากโรคหัวใจและหลอดเลือดเท่ากับ 63.8 คนต่อประชากรหนึ่งแสนคน แม้ว่าประเทศไทยจะมีสิทธิประกันสังคม ม.33 ม.39 และบัตรทองที่ให้ตรวจสุขภาพประจำปีฟรี ซึ่งสามารถช่วยให้ตรวจพบโรคหัวใจตั้งแต่ระยะแรก ๆ แต่อัตราการเสียชีวิตจากโรคหัวใจและหลอดเลือดในประเทศไทยยังคงลดลงเพียงเล็กน้อย (กองโรคไม่ติดต่อ, 2566) ซึ่งอาจเกิดจากความแออัดของโรงพยาบาลรัฐ

ความแออัดของโรงพยาบาลรัฐเป็นปัญหาเรื้อรังของระบบสาธารณสุขไทย ซึ่งส่งผลให้ผู้ป่วยต้องรอพบแพทย์นาน บุคลากรทางการแพทย์เหนื่อยล้าจากการทำงานจนกระทบต่อประสิทธิภาพในการรักษา รวมถึงก่อให้เกิดความตั้งใจไม่ใช้สิทธิตรวจสุขภาพประจำปีฟรี (ชลธร วงศ์รัศมี, 2561) ฉะนั้นเพื่อลดอัตราการเสียชีวิตจากโรคหัวใจและหลอดเลือดในประเทศไทย การซื้อเครื่องวัดความดันโลหิตแบบดิจิตอลมาใช้เองที่บ้าน แล้วใช้ผลจากเครื่องวัดความดันโลหิตดังกล่าวมาประกอบการตัดสินใจในการเข้ารับการบริการที่โรงพยาบาล และการประเมินการรักษาของแพทย์จึงอาจช่วยลดอัตราการเสียชีวิตจากโรคหัวใจและหลอดเลือดได้

เครื่องวัดความดันโลหิตแบบดิจิตอลสามารถวัดได้ทั้งความดันโลหิตและอัตราการเต้นของหัวใจ โดยความดันโลหิตสูง (ค่าความดันซิสโตลีมากกว่าหรือเท่ากับ 140 มิลลิเมตรปรอท และค่าความดันไดแอสโตลีมากกว่าหรือเท่ากับ 90 มิลลิเมตรปรอท ในสภาวะพัก) เป็นปัจจัยที่เสี่ยงทำให้เกิดโรคหัวใจ (โรงพยาบาลรามคำแหง, 2567) ขณะที่หากอัตราการเต้นของหัวใจเกิน 100 ครั้งต่อนาทีในสภาวะพักอาจเกิดจากการเป็นโรคหัวใจ โดยก่อนวัดความดันโลหิตควรงดบุหรี่ แอลกอฮอล์ และคาเฟอีนอย่างน้อย 30 นาที รวมถึงนั่งพักอย่างน้อย 5 นาที (โรงพยาบาลสินแพทย์ รามอินทรา, 2564)

ดังนั้น เพื่อให้ทราบว่าความดันโลหิต อัตราการเต้นของหัวใจและปัจจัยอื่น ๆ เป็นสาเหตุของโรคหัวใจมากน้อยเพียงใด แบบจำลองแบ่งแยกประเภท (Classification Model) จึงเป็นเครื่องมือที่สามารถนำมาช่วยระบุน้ำหนักของปัจจัยต่าง ๆ ที่ใช้ในการคาดการณ์ว่าเป็นหรือไม่เป็นผู้ป่วยโรคหัวใจ โดยงานวิจัยนี้สามารถนำไปเป็นข้อมูลประกอบการตัดสินใจซื้อเครื่องดันความดันโลหิตแบบดิจิตอล เพื่อประกอบการตัดสินใจในการเข้ารับการบริการที่โรงพยาบาล

1.2 วัตถุประสงค์

เพื่อพัฒนาแบบจำลองแบ่งแยกประเภท (Classification Model) ที่ดีที่สุด และระบุน้ำหนักของความดันโลหิต อัตราการเต้นของหัวใจ และปัจจัยอื่น ๆ ที่ใช้ในการคาดการณ์ว่าเป็นหรือไม่เป็นผู้ป่วยโรคหัวใจ

1.3 ขอบเขตงานศึกษา

1. แบบจำลองแบ่งแยกประเภท (Classification Model) ที่ใช้ ได้แก่ แบบจำลองต้นไม้ตัดสินใจ (Decision Tree) และแบบจำลองป่าสุ่ม (Random Forest)

2. เกณฑ์ที่ใช้ในการคัดเลือกแบบจำลอง คือ Accuracy Score เพื่อวัดความแม่นยำในการจำแนกประเภทในเบื้องต้นอย่างไม่เจาะจงกลุ่มตัวอย่างใด ๆ

3. ข้อมูลของกลุ่มตัวอย่างที่ใช้เป็นข้อมูลในการสร้างแบบจำลอง ได้จากการหาข้อมูลโดยการสืบค้นบนอินเทอร์เน็ต คือ ชุดข้อมูล “Heart Disease Dataset” ซึ่งประกอบด้วยข้อมูลจำนวน 1190 ตัวอย่าง

4. โปรแกรมที่ใช้ในการวิเคราะห์ข้อมูลและพัฒนาแบบจำลอง คือ Python

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 Machine Learning from SKLearn

Machine Learning เป็นสาขาหนึ่งของปัญญาประดิษฐ์ (AI) และวิทยาการคอมพิวเตอร์ ที่มุ่งเน้นไปที่การใช้ข้อมูลและอัลกอริทึม (Algorithm) เพื่อให้ AI เลียนแบบวิธีที่มนุษย์เรียนรู้ และค่อยๆ เพิ่มความแม่นยำขึ้น

ระบบการเรียนรู้ของอัลกอริทึมใน Machine Learning สามารถแบ่งออกได้เป็นสามส่วนตามหลักของ UC Berkeley

1. กระบวนการตัดสินใจ (Decision Process) โดยทั่วไปแล้ว อัลกอริทึมใน Machine Learning จะใช้ในการทำนายหรือจำแนกประเภท จากข้อมูลนำเข้า (input) บางอย่าง ซึ่งจะมีคำตอบ (Label) หรือไม่มีคำตอบก็ได้ และอัลกอริทึมจะสร้างค่าประมาณ โดยเชื่อมโยงกับรูปแบบจากข้อมูลทั้งหมดที่มี

2. ฟังก์ชันข้อผิดพลาด (Error Function) ฟังก์ชันข้อผิดพลาดจะประเมินการทำนายของแบบจำลอง หากพบข้อผิดพลาดขึ้น ฟังก์ชันข้อผิดพลาดสามารถทำการเปรียบเทียบเพื่อประเมินความแม่นยำของแบบจำลองได้

3. กระบวนการเพิ่มประสิทธิภาพโมเดล (Model Optimization Process) หากโมเดลสามารถรองรับกับข้อมูลในชุดการฝึก (Train Data) ได้ดีขึ้น น้ำหนักในการคำนวนของข้อมูลตัวอย่างที่ทราบจะถูกปรับเพื่อลดความแตกต่างกับการประมาณแบบจำลอง อัลกอริทึมจะทำซ้ำกระบวนการประเมินและเพิ่มประสิทธิภาพซ้ำ โดยจะอัปเดตน้ำหนักอัตโนมัติจนกว่าจะถึงเกณฑ์ความแม่นยำที่ตั้งไว้

SKLearn หรือ **Scikit-learn** เป็นไลบรารีแบบ open-source (ชุดของโค้ดโปรแกรมที่นักพัฒนาได้เขียนขึ้นมาและแจกจ่ายให้ใช้งานฟรีโดยไม่มีค่าใช้จ่าย อีกทั้งยังเปิดเผยซอร์สโค้ด (source code) ให้ผู้ใช้งานสามารถดู แก้ไข และนำไปพัฒนาต่อได้ตามความต้องการ) ไลบรารีนี้เกี่ยวข้องกับ Machine Learning สำหรับภาษา Python ซึ่งพัฒนาโดย **David Cournapeau** โดยในปัจจุบันนับว่าเป็นที่นิยมอย่างมากสำหรับการเรียนรู้และพัฒนาด้าน Machine Learning (UC Berkeley, 2565)

2.1.1 การ Train และ Test ใน Machine Learning from SKLearn ด้วยภาษา Python

การ Train และ Test ใน Machine Learning เป็นกระบวนการที่สำคัญในการสร้างแบบจำลอง และประเมินประสิทธิภาพของแบบจำลองที่ได้สร้างขึ้นมา โดยการ Train แบบจำลอง จะ

เป็นการสร้างแบบจำลองจากชุดข้อมูล (dataset) ที่มีคำตอบ (label) ที่เป็นจริง โดยการ Train แบบจำลองจะใช้ชุดข้อมูลที่มีคำตอบเพื่อเรียนรู้ความสัมพันธ์ระหว่างตัวแปร (features) และคำตอบ ซึ่งแบบจำลองจะนำไปใช้ในการทำนายคำตอบของข้อมูลใหม่ที่ไม่เคยเห็นมาก่อน หลังจาก Train แบบจำลองแล้ว จะทำการประเมินประสิทธิภาพของแบบจำลองด้วยการใช้ชุดข้อมูลที่ไม่มีคำตอบเพื่อทำนายคำตอบ ซึ่งเรียกว่าการ Test แบบจำลอง โดยการ Test แบบจำลองจะทำการนำชุดข้อมูลที่ไม่มีคำตอบมาใช้ทำนายคำตอบ แล้วเปรียบเทียบคำตอบจริงกับคำตอบที่แบบจำลองทำนายได้ เพื่อวัดประสิทธิภาพของแบบจำลองว่าทำนายได้ถูกต้องแค่ไหน โดยการ Train และ Test ใน Machine Learning สามารถทำได้หลายวิธี แต่วิธีที่นิยมใช้กันคือการแบ่งชุดข้อมูลเป็นสองส่วน ซึ่งส่วนหนึ่งจะใช้สำหรับ Train แบบจำลอง และส่วนที่เหลือจะใช้สำหรับ Test แบบจำลอง วิธีนี้เรียกว่า Holdout Validation (กิตติมศักดิ์ ในจิต, ม.ป.ป.)

2.1.2 วิธีการ Train ใน Machine Learning

1. สร้างโมเดล (Model) โดยใช้ Algorithm หรือวิธีการใด ๆ ที่เหมาะสมกับปัญหา และระบุพารามิเตอร์ที่เหมาะสมต่อการ Train

2. สร้างชุดข้อมูลสำหรับ Train โมเดล โดยแบ่งข้อมูลออกเป็นสองส่วนคือ Training Set และ Validation Set

3. นำ Training Set มา Train ในโมเดลโดยใช้ Algorithm หรือวิธีการที่เลือกไว้ แล้วปรับพารามิเตอร์ต่าง ๆ ให้เหมาะสมกับข้อมูลที่เลือก

4. ใช้ Validation Set เพื่อวัดประสิทธิภาพของโมเดลที่ Train โดยการประเมินผลของโมเดลที่ Train ด้วย Metrics ที่เลือกไว้ เช่น Accuracy, F1 Score, Precision, Recall หรืออื่น ๆ

5. ปรับแก้ไขโมเดลหรือพารามิเตอร์ต่าง ๆ ตามผลการประเมินจาก Validation Set จนกว่าโมเดลจะมีประสิทธิภาพดีที่สุด

6. นำโมเดลที่ Train ด้วย Training Set และ Validation Set ที่ปรับแก้ไขแล้ว มาทดสอบกับ Test Set โดยวัดผลด้วย Metrics ที่เลือกไว้ เพื่อประเมินประสิทธิภาพของโมเดลโดยทั่วไป

2.1.3 วิธีการ Test ใน Machine Learning

การ Test ใน Machine Learning เป็นกระบวนการที่ใช้ข้อมูลทดสอบ (test data) เพื่อประเมินประสิทธิภาพของโมเดลที่ได้ทำการสร้างขึ้นจากข้อมูลฝึก (training data)

โดยทั่วไปแล้ว ข้อมูลทดสอบจะไม่ถูกใช้ในกระบวนการ Train โมเดล และในการทดสอบโมเดลส่วนมากมักจะใช้ฟังก์ชัน predict โดยการทดสอบโมเดลทั่วไปมีขั้นตอนดังนี้

1. โหลดข้อมูลทดสอบ (test data) โดยแยก feature และ label ออกจากกัน

2. นำโมเดลที่ได้ Train ไว้แล้ว มาใช้กับข้อมูลทดสอบเพื่อทำนายผลลัพธ์

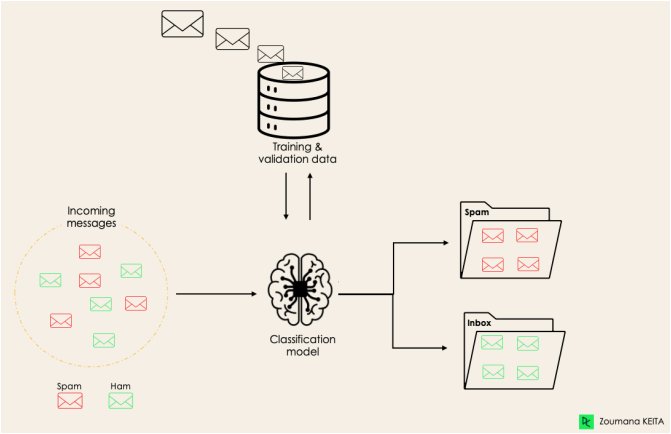
3. ประเมินประสิทธิภาพของโมเดลด้วย metrics ต่างๆ เช่น accuracy, precision, recall, F1-score

4. ทำการทดสอบไปเรื่อย ๆ จนกว่าประสิทธิภาพของโมเดลจะถูกต้องตามที่ต้องการ

2.2 Classification Model

Classification Model เป็นโมเดลที่มีการเรียนรู้แบบมีผู้สอนใน Machine Learning โดยที่โมเดลจะพยายามคาดเดาคำตอบที่ถูกต้องของข้อมูลนำเข้า (Input) ที่กำหนด ในการจำแนกประเภท โมเดลจะได้รับการฝึกโดยใช้ข้อมูลการฝึก (Train Data) จากนั้นจะมีการประเมินกับข้อมูลทดสอบ (Test Data) ก่อนที่จะนำไปใช้ในการทำนายข้อมูล ตัวอย่างเช่น อัลกอริทึมสามารถเรียนรู้ที่จะคาดเดาว่าอีเมลที่ระบุนั้นเป็นสแปมหรือไม่ ดังตัวอย่างจากรูปที่ 2.1

รูปที่ 2.1 หลักการทำงานของ Classification Model (Datacamp, 2565)



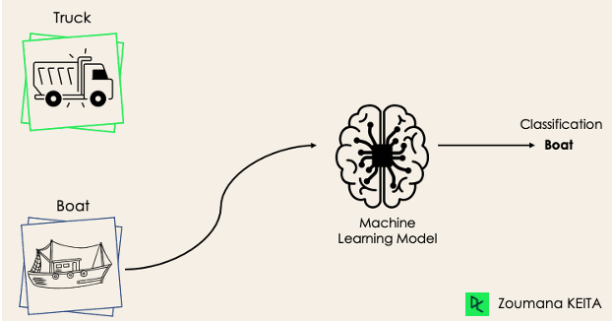
Classification Model สามารถแบ่งออกได้เป็น 4 ประเภท ดังนี้

1. Binary Classification (การจำแนกประเภทแบบไบนารี)

เป้าหมายของการจำแนกแบบนี้คือการจัดประเภทข้อมูลนำเข้าออกเป็นสองประเภท ซึ่งเกิดขึ้นแยกกัน โดยผลลัพธ์ของโมเดลนี้ จะมีเพียง 2 คำตอบเท่านั้น เช่น จริงหรือเท็จ, บวกหรือลบ, 0

หรือ 1, สแปมหรือไม่ใช่สแปม ฯลฯ โดยคำตอบจะขึ้นอยู่กับปัญหาที่กำลังพบอยู่ ตัวอย่างเช่น เราอาจต้องการตรวจสอบว่ารูปภาพที่ระบุเป็นรถบรรทุกหรือเรือ

รูปที่ 2.2 Binary Classification (Datacamp, 2565)



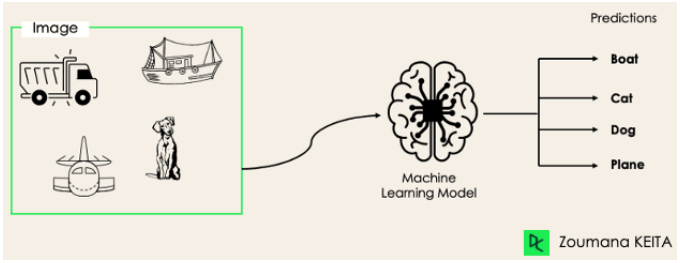
2. Multi-Class Classification (การจำแนกประเภทแบบหลายคลาส)

เป้าหมายของการจำแนกแบบนี้คือการจัดประเภทข้อมูลนำเข้าให้มีคำตอบมากกว่า 2 คำตอบขึ้นไป โดยที่อัลกอรึทึมของ Binary Classification ส่วนมากสามารถใช้กับ Multi-Class Classification ได้

3. Multi-Label Classification (การจำแนกประเภทแบบหลายคำตอบ)

ในการจำแนกประเภทแบบหลายคำตอบ โมเดลจะพยายามทำนายคำตอบของแต่ละข้อมูลนำเข้าให้มีจำนวนเป็น 0 หรือมากกว่า โดยในกรณีนี้ข้อมูลนำเข้าสามารถนำเข้าร่วมกันได้ เนื่องจากตัวอย่างอินพุตสามารถมีคำตอบได้มากกว่าหนึ่งอย่าง ยกตัวอย่างเช่น รูปภาพหนึ่งสามารถมีวัตถุได้หลายชิ้น ดังรูปที่ 2.3 แบบจำลองจะสามารถคาดการณ์ได้ว่าในรูปภาพประกอบด้วยเครื่องบิน เรือ รถบรรทุก และสุนัข

รูปที่ 2.3 Multi=Label Classification (Datacamp, 2565)

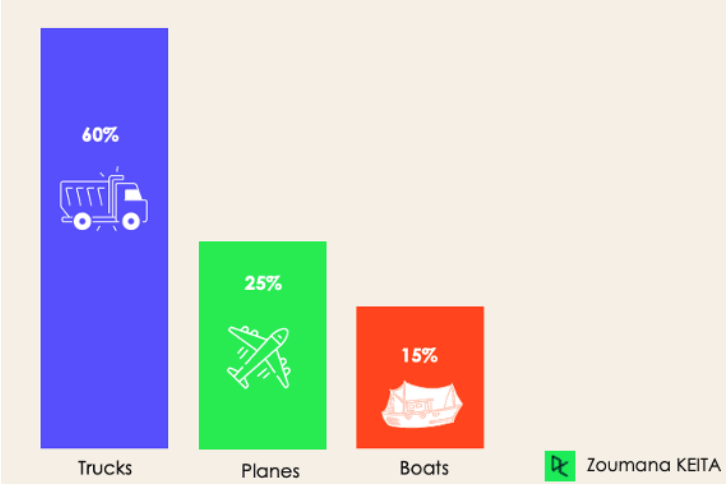


4. Imbalanced Classification (การจำแนกประเภทแบบไม่สมดุล)

สำหรับการจำแนกประเภทแบบไม่สมดุล จำนวนตัวอย่างจะกระจายไม่เท่ากันในแต่ละคลาส ซึ่งหมายความว่าเราสามารถมีคลาสได้มากกว่าหนึ่งคลาสในข้อมูลการฝึก (Train Data) โดยจะลอง

พิจารณาสถานการณ์การจำแนกประเภท 3 ระดับต่อไปนี้ โดยที่ข้อมูลการฝึกประกอบด้วย รถบรรทุก 60%, เครื่องบิน 25% และ เรือ 15%

รูปที่ 2.4 Imbalanced Classification (Datacamp, 2565)



ปัญหาการการจำแนกประเภทแบบไม่สมดุลอาจเกิดขึ้นได้ในสถานการณ์ต่อไปนี้ การตรวจจับธุรกรรมที่เป็นการฉ้อโกงในอุตสาหกรรมการเงิน, การวินิจฉัยโรคที่หายาก หรือการวิเคราะห์การเลิกใช้งานของลูกค้า การใช้แบบจำลองการทำนายแบบเดิม ๆ ไม่มีประสิทธิภาพพอเมื่อต้องรับมือกับชุดข้อมูลที่ไม่สมดุล เนื่องจากอาจเกิดอคติต่อการทำนายคลาสที่มีจำนวนการสังเกตสูงสุด และพิจารณาว่าโมเดลที่มีจำนวนน้อยกว่านั้นเป็นการรบกวน แต่ไม่ได้หมายความว่าปัญหานี้จะไม่สามารถถูกแก้ได้ เรายังสามารถใช้หลายวิธีเพื่อจัดการกับปัญหาความไม่สมดุลในชุดข้อมูล โดยวิธีการที่นิยมใช้บ่อยที่สุด ได้แก่ การสุ่มตัวอย่างโดยคำนึงถึงต้นทุน

2.3 Regression

Regression เป็นการศึกษาความสัมพันธ์ของตัวแปร 2 ตัวขึ้นไป โดยส่วนใหญ่จะใช้ตัวแปร x, y แทนตัวแปรที่ต้องการศึกษา ซึ่งตัวแปรดังกล่าวจะมีความสัมพันธ์ในลักษณะที่เกี่ยวข้องซึ่งกันและกัน โดยในภาษาไทย Regression จะแปลว่าการถดถอย ซึ่งในทาง Machine Learning แล้วจะไม่สามารถแปลว่าการถดถอยโดยตรงได้ เนื่องจากในทาง Machine Learning จะสามารถหมายถึงการเพิ่มขึ้นได้อีกด้วย ดังนั้นการใช้ทับศัพท์ว่า Regression ไปเลยจะทำให้เกิดความเข้าใจที่ตรงกันมากกว่า

Regression คือการนำเอาข้อมูลที่เก็บไว้ในอดีตมาทำนายแนวโน้มข้อมูลที่จะเกิดขึ้นในอนาคต โดยใช้รูปแบบสมการเชิงเส้น (Linear) โดยใช้วิธีการหาความสัมพันธ์ของตัวแปร 2 ตัวขึ้นไป โดยเราจะต้องระบุให้ชัดเจนว่าตัวแปรใด คือ ตัวแปรอิสระ/ตัวแปรต้น (กำหนดเป็นค่า X) และตัวแปรใด คือ ตัวแปรตาม (กำหนดให้เป็น y)

2.4 Linear Regression Model

Linear Regression เกิดจากการรวมกันของคำว่า Linear ที่แปลว่าเชิงเส้น และ Regression ที่แปลว่าการถดถอย ซึ่ง Linear Regression จะตีความหมายได้ว่าความสัมพันธ์ของตัวแปรหรือสิ่งที่เรากำลังสนใจ ซึ่งจะถูกใช้กับการคำนวณค่าที่เป็นตัวเลข เพื่อหาความสัมพันธ์หรือทำนายข้อมูลต่าง ๆ Linear Regression ถือว่าเป็น Machine Learning ประเภท Supervised Learning (การเรียนรู้แบบมีผู้สอน) โดยที่เราจะต้องใส่ชุดข้อมูลเข้าไปให้โปรแกรมเรียนรู้ก่อน และโปรแกรมจึงจะนำตัวแปรต้นและตัวแปรตามไปคำนวณด้วยสถิติทางคณิตศาสตร์ แล้วโปรแกรมจะคืนข้อมูลกลับมาเป็นตัวเลข

ในสมการ Linear Regression Model เราจะให้สมการความสัมพันธ์ของตัวแปรเป็นดังนี้

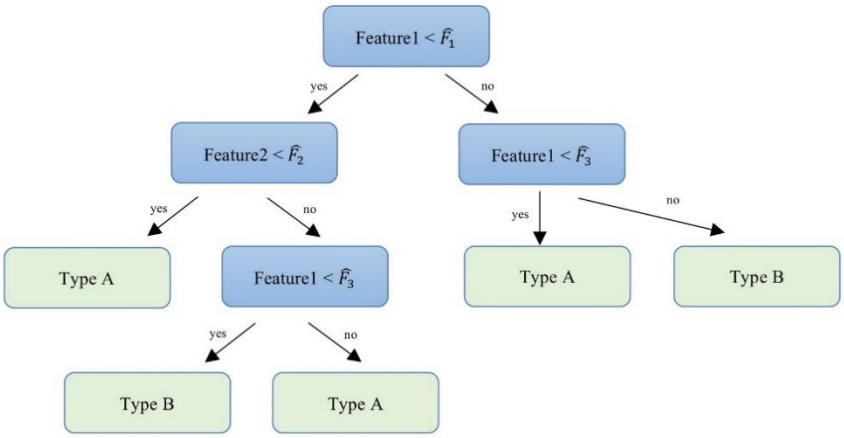
$$y_i = w_0 + w_1 x_i + \epsilon_i$$

โดยที่ w_0 คือจุดตัดของแกน Y และ w_1 คือความชันของสมการ ϵ_i คือความผิดพลาดในการประมาณค่า หรือความแปรผันในการประมาณค่า และ w_0 , w_1 เป็นค่าพารามิเตอร์ของโมเดลที่เราต้องการให้เรียนรู้ (Phanpaporn, ม.ป.ป.)

2.5 Decision Tree Model

แบบจำลองต้นไม้ตัดสินใจ เป็นแบบจำลองที่ใช้วิธีการจำแนกประเภทข้อมูลด้วยการใช้เกณฑ์คัดแยกเป็นขั้น ๆ ต่อเนื่องไปเรื่อย ๆ โดยบางขั้นจะสามารถนำตัวแปรคุณลักษณะที่เคยถูกใช้ไปแล้ว นำกลับมาเป็นเกณฑ์ในการจำแนกได้อีก โดยการจำแนกเป็นขั้น ๆ แบบนี้ หากทำเป็นแผนภูมิแล้วจะมองเห็นเป็นแผนภาพคล้ายกับต้นไม้ได้ โดยด้านบนจะเป็นราก และแตกกิ่งออกไปเรื่อย ๆ ตามขั้นตอนที่ตัดสินใจ

รูปที่ 2.5 Decision Tree Model (Teerawut Sripinit, 2567)



จะเห็นได้ว่ารูปที่ 2.5 แสดงถึงการตัดสินใจครั้งที่ 1 เป็นเกณฑ์ โดยค่าที่เหมาะสมสำหรับข้อมูลไว้ที่ $\hat{F}_1$ โดยที่หากข้อมูลมีค่าน้อยกว่าที่กำหนดไว้ ค่าก็จะถูกจำแนกไปทางด้านซ้าย แต่หากมากกว่า ก็จะถูกจำแนกไปด้านขวา และจะจำแนกไปเรื่อย ๆ จนกว่าข้อมูลที่ถูกแยกจะอยู่ในกลุ่มย่อยอย่างชัดเจน หรือจนกว่าลำดับของการตัดสินใจจะสิ้นสุดตามที่กำหนดไว้ ซึ่งวิธีที่เราจะคัดเลือกตัวแปรคุณลักษณะ และค่าของคุณลักษณะที่จะนำมาใช้ให้มีประสิทธิภาพ จะมีวิธีที่นิยมนำมาพิจารณาอยู่ 2 วิธี คือ Information gain และ Gini impurity แต่ทั้ง 2 วิธีนี้ต่างก็มีเป้าหมายแบบเดียวกันคือจัดการข้อมูลที่คัดกรองแล้วให้เป็นระเบียบมากขึ้น

2.5.1 ประเภทของ Decision Tree

2.5.1.1 Regression Tree

Regression Tree เป็น Decision Tree ที่ใช้เพื่อใช้หาความสัมพันธ์ของตัวแปรต่าง ๆ โดยอิงจากค่าของ Residual sum of squares (RSS) เพื่อให้หาจุดที่ดีที่สุดในการแบ่งข้อมูลด้วยการทำให้ RSS มีค่าต่ำที่สุด ซึ่งมีสมการดังนี้

$$RSS = \sum_{n=1}^{N} \sum_{i \in R} (y_i - \hat{y}_{Rn})^2$$

โดยที่ Rn = แต่ละกลุ่มของ กลุ่มตัวอย่างที่ถูกแบ่งออกมา เป็นทั้งหมด N กลุ่ม และ yi = target variable $\hat{y}_{Rn}$ = ค่าที่คาดหวังในแต่ละกลุ่ม โดยคำนวณมาจากค่า mean ของ target variable ในกลุ่มนั้นๆ

Mean Square Error เป็นการคำนวณโอกาสที่จะผิดพลาด โดยมีสมการดังนี้

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

จากสมการข้างต้น y_i จะหมายถึงค่าคาดหมาย และ $\hat{y}_i$ จะหมายถึงค่าที่เป็นความจริง (Actual)

2.5.1.2 Classification Tree

หลักการของ Classification Tree เหมือนกับ Regression Tree แต่ต่างกันตรงที่ค่า Cost Function จาก Regression ที่ใช้ RSS เปลี่ยนเป็น Gini impurity หรือ Entropy เพื่อการแก้ไขปัญหา Classification ให้เหมาะสม

Gini impurity เป็นการวัดความไม่บริสุทธิ์ของ class ในแต่ละกลุ่มข้อมูลที่แบ่งแยกย่อยออกมา สำหรับปัญหา Binary Classification (การจำแนกประเภทแบบไบนารี) ที่มี ตัวแปรเป้าหมาย เป็น 0 หรือ 1 การจำแนกที่ดี ควรจะแบ่งข้อมูลออกมาได้เป็น 2 หากสามารถแยก class 0 กับ class 1 ออกมาได้ชัดเจนในแต่ละกลุ่ม ยิ่งสามารถแบ่งแยก class ของ target variable ออกมาได้ดี ค่า Gini impurity ก็จะยิ่งมีค่าน้อยลง

Gini impurity มีสูตรในการคำนวณดังนี้

$$Gini\ impurity = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

จากสมการข้างต้น K หมายถึงจำนวน Class ของตัวแปรเป้าหมาย และค่าของ $\hat{p}_{mk}$ จะหมายถึงสัดส่วนของ Class K ในแต่ละกลุ่ม โดยในสมการจะเป็นการนำเอาผลรวมของค่าความน่าจะเป็นของเหตุการณ์ที่สนใจคูณกับความน่าจะเป็นที่เหลือ

Entropy เป็นการวัดความไม่แน่นอน (randomness) ของข้อมูล สำหรับการทำ model classification เราต้องการทำนาย class ของตัวแปรเป้าหมายให้แม่นยำมากที่สุด หมายความว่า เราต้องการลดความไม่แน่นอน หรือ randomness ให้เหลือน้อยที่สุด ซึ่งจะใช้การพยายามแยก class ของตัวแปรเป้าหมาย ให้มีสัดส่วนใน class หนึ่งมากที่สุด เพื่อเพิ่มความแม่นยำในการทำนาย

Entropy มีสูตรในการคำนวณดังนี้

$$Entropy = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$$

สมการของ Entropy มีความคล้ายคลึงกับ Gini impurity แต่จะต่างกันตรงที่เปลี่ยนจากความน่าจะเป็นที่เหลือ เป็น log ของความน่าจะเป็นที่สนใจแทน

2.5.2 วิธีการใช้งาน Decision Tree ใน Python

Classification tree และ Regression tree ต่างก็ต้องถูกนำเข้าผ่าน SKLearn ใน Python โดยที่จะมีพารามิเตอร์ที่สำคัญดังนี้

1. criterion คือ cost function ที่เราจะใช้สำหรับ regression tree จะมีค่า default เป็น mean square error ส่วน classification tree จะ มีค่า default เป็น gini ซึ่งเราสามารถที่จะเลือกเป็น entropy ได้

2. max depth คือ จำนวนชั้นที่มากที่สุดของกลุ่มที่จะทำการจำแนก ซึ่งเราจะกำหนดค่าของ max depth ไม่ให้มากเกินไป เพื่อป้องกันปัญหา Overfitting (Model ที่ได้จะสามารถเรียนรู้ข้อมูลจาก Training Data set ได้ดีมาก แต่ไม่สามารถนำไปใช้กับข้อมูลที่ไม่เคยพบมาก่อนได้ดี)

3. min samples leaf คือ จำนวนการสำรวจขั้นต่ำที่จะให้อยู่ในกลุ่ม ถ้าหากเกิดค่าดังกล่าวน้อยจนเกินไป ก็อาจทำให้เกิดปัญหา overfitting ได้

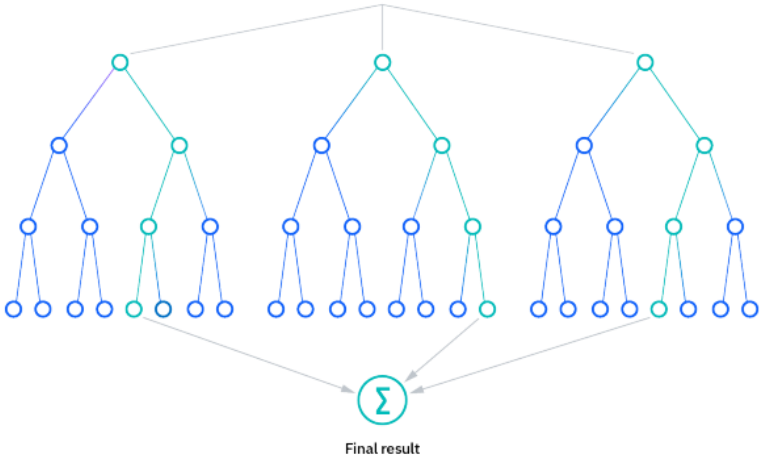
2.6 Random Forest Model

แบบจำลองป่าไม้สุ่มเป็นอัลกอริทึมใน Machine ที่นิยมใช้อย่างมาก ซึ่งถูกคิดค้นโดย Leo Breiman และ Adele Cutler โดยจะเป็นการรวมผลลัพธ์ (Output) ของ Decision Tree หลาย ๆ ต้นเพื่อให้ได้ผลลัพธ์เดียว เพื่อให้เกิดความสะดวกและมีความยืดหยุ่นในการใช้งาน และยังสามารถจัดการกับปัญหา Classification และ Regression ได้ทั้ง 2 ประเภท ทำให้แบบจำลองนี้เป็นที่ยอมรับกันอย่างแพร่หลาย

แบบจำลองป่าไม้สุ่มประกอบด้วย Decision Tree หลายต้น ซึ่ง Decision Tree จะเริ่มต้นด้วยคำถามพื้นฐาน เช่น ฉันควรไปทะเลไหม จากนั้นเรายังสามารถถามคำถามหลายชุดเพื่อหาคำตอบได้ เช่น ใช้เวลาเดินทางนานหรือไม่ หรือจะมีฝนตกบริเวณชายหาดหรือไม่ คำถามเหล่านี้จะประกอบขึ้นเป็นกลุ่มของ Decision Tree ซึ่งทำหน้าที่เป็นช่องทางในการแบ่งข้อมูล คำถามแต่ละข้อจะช่วยให้สามารถตัดสินใจขั้นสุดท้ายได้ ข้อมูลที่เข้าเกณฑ์จะถูกจำแนกไปที่ Yes และข้อมูลที่ไม่เข้าเกณฑ์จะถูกจำแนกไปที่ No Decision Tree จะพยายามหาการจำแนกที่ดีที่สุดเพื่อคัดแยกข้อมูล และโดยทั่วไปแล้ว Decision Tree จะได้รับการฝึกผ่านอัลกอริทึม Classification and Regression Tree เช่น Gini ของข้อมูลที่ได้รับ หรือ MSE ซึ่งสามารถใช้เพื่อประเมินคุณภาพของการจำแนกได้

อัลกอริทึมของแบบจำลองป่าไม้สุ่มมีไฮเปอร์พารามิเตอร์หลักสามตัว ซึ่งจำเป็นต้องตั้งค่าก่อนการฝึก ซึ่งรวมถึงขนาดของกลุ่ม จำนวนแผนผัง และจำนวนคุณลักษณะที่สุ่มตัวอย่าง จากนั้น Random forest classifier จึงจะสามารถแก้ปัญหา Regression และ Classification ได้

รูปที่ 2.6 แบบจำลอง Random Forest (IBM, ม.ป.ป.)



Random Forest Model มีข้อดีคือ ลดความเสี่ยงของการ Overfitting เนื่องจาก Decision Tree มีความเสี่ยงในการ Overfitting มากเกินไป แต่เมื่อมี Decision Tree จำนวนมากใน Random Forest การจำแนกประเภทจะไม่ Overfit กับโมเดลมากเกินไป เนื่องจากการหาค่าเฉลี่ยของ Decision Tree ที่ไม่เกี่ยวข้องกันจะช่วยลดความแปรปรวนโดยรวมและข้อผิดพลาดในการทำนายได้ดีมากยิ่งขึ้น และ Random Forest ยังมีความยืดหยุ่นสูง เนื่องจาก Random Forest สามารถจัดการทั้งปัญหา Regression และ Classification ได้ด้วยความแม่นยำสูง จึงเป็นวิธีที่ได้รับความนิยมในหมู่นักวิทยาศาสตร์ข้อมูลอีกด้วย อย่างไรก็ตาม Random Forest ก็ยังมีความท้าทายที่ต้องเผชิญอยู่ เช่น การใช้เวลานาน เนื่องจากอัลกอริทึมของ Random Forest สามารถจัดการชุดข้อมูลขนาดใหญ่ได้ และสามารถให้การคาดการณ์ที่แม่นยำมากขึ้น แต่อาจประมวลผลข้อมูลได้ช้า เนื่องจากเป็นการประมวลผลข้อมูลสำหรับ Decision Tree แต่ละแบบ และยังต้องการทรัพยากรมากขึ้น เพราะว่า Random Forest ต้องประมวลผลชุดข้อมูลที่มีขนาดใหญ่ขึ้น ทำให้มีความต้องการทรัพยากรมากขึ้นในการจัดเก็บข้อมูลนั้น ๆ อีกทั้งยังมีความซับซ้อน เพราะการทำนาย Decision Tree เดียวจะตีความได้ง่ายกว่าเมื่อเปรียบเทียบกับ Random Forest (Witchapong Daroontham, 2561)

2.7 Confusion Matrix

ใน Machine Learning การจำแนกประเภท คือกระบวนการจัดหมวดหมู่ชุดข้อมูลที่กำหนดออกเป็นหมวดหมู่ต่าง ๆ ซึ่งการวัดประสิทธิภาพของโมเดลการจำแนกประเภทต่าง ๆ เราจะใช้เมทริกซ์วัดประสิทธิภาพในการวัด โดย Confusion Matrix คือเมทริกซ์ที่จะสรุปประสิทธิภาพของโมเดล Machine Learning กับชุดข้อมูลทดสอบ (Test Data) เป็นวิธีการแสดงว่าข้อมูลแม่นยำหรือไม่ตามการคาดการณ์ของแบบจำลอง โดยมักจะถูกใช้เพื่อวัดประสิทธิภาพของแบบจำลองการจำแนกประเภท ซึ่งมีจุดมุ่งหมายเพื่อทำนายคำตอบ (Label) สำหรับแต่ละข้อมูลนำเข้า

เมทริกซ์จะแสดงจำนวนกรณีที่สร้างโดยโมเดลบนข้อมูลทดสอบ โดยจะมีทั้งหมด 4 กรณี

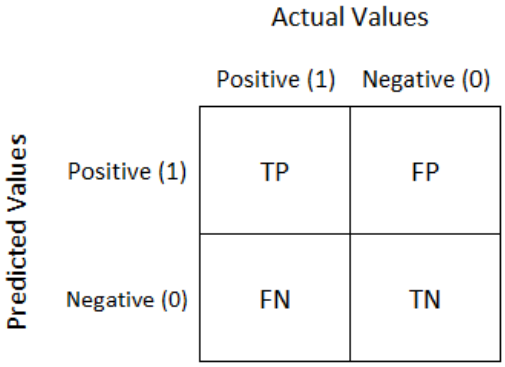
1. True positives (TP) เกิดขึ้นเมื่อโปรแกรมทำนายว่าจริง และมีค่าเป็นจริง

2. True negatives (TN) เกิดขึ้นเมื่อโปรแกรมทำนายว่าไม่จริง และมีค่าเป็นไม่จริง

3. False positives (FP) เกิดขึ้นเมื่อโปรแกรมทำนายว่าจริง และมีค่าเป็นไม่จริง

4. False negatives (FN) เกิดขึ้นเมื่อโปรแกรมทำนายว่าไม่จริง และมีค่าเป็นจริง

รูปที่ 2.7 Confusion Matrix (Medium, 2019)



หน่วยวัดใน Confusion Matrix

2.7.1 Accuracy

ความแม่นยำใช้ในการวัดประสิทธิภาพของแบบจำลอง เป็นอัตราส่วนของกรณีที่ถูกต้องทั้งหมดต่อกรณีทั้งหมด

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

2.7.2 Precision

เป็นการวัดความแม่นยำของการคาดการณ์เชิงบวกของแบบจำลอง มักถูกกำหนดให้เป็นอัตราส่วนของการทำนายเชิงบวกที่แท้จริงต่อจำนวนการคาดการณ์เชิงบวกทั้งหมดที่ทำโดยแบบจำลอง

$$Precision = \frac{TP}{TP + FP}$$

2.7.3 Recall

เป็นการวัดประสิทธิภาพของแบบจำลองในการจำแนกประเภท โดยจะระบุกรณีที่เกี่ยวข้องทั้งหมดจากชุดข้อมูล เป็นอัตราส่วนของจำนวนกรณี True Positive (TP) ต่อผลรวมของกรณี True Positive และ False Negative (FN)

$$Recall = \frac{TP}{TP + FN}$$

2.7.4 F1-Score

จะใช้เพื่อประเมินประสิทธิภาพโดยรวมของแบบจำลองการจำแนกประเภท

$$F1 - Score = \frac{2(Precision * Recall)}{Precision + Recall}$$

บทที่ 3

วิธีการดำเนินการวิจัย

3.1 การรวบรวมข้อมูลเพื่อใช้สร้างแบบจำลอง

ผู้ศึกษาได้ข้อมูลข้อมูลของกลุ่มตัวอย่างที่ใช้ในการสร้างแบบจำนวน จากการหาข้อมูลโดยการสืบค้นบนอินเทอร์เน็ต คือ ชุดข้อมูล “Heart Disease Dataset” ซึ่งประกอบด้วยข้อมูลจำนวน 1190 ตัวอย่าง โดยมี 12 ตัวแปร

3.2 การกำหนดตัวแปร

3.2.1 การกำหนดตัวแปรอิสระ

ผู้ศึกษาได้นำตัวแปรอิสระจากการเก็บข้อมูลผู้ที่เป็นโรคหัวใจและไม่เป็นโรคหัวใจ ซึ่งเป็นข้อมูลที่ค้นพบบนอินเทอร์เน็ต โดยมีตัวแปรอิสระทั้งหมดจำนวน 11 ตัวแปร ได้แก่ อายุ (age) เพศ (sex) ลักษณะอาการเจ็บหน้าอก (chest pain type) ความดันโลหิตขณะพัก (resting bp s) ระดับของคอเลสเตอรอลในเลือด (cholesterol) ระดับน้ำตาลในเลือดในช่วง 2-3 วันที่ผ่านมา โดยงดอาหาร 8 ชั่วโมงก่อนตรวจ (fasting blood sugar) คลื่นไฟฟ้าหัวใจขณะพัก (resting ecg) อัตราการเต้นของหัวใจสูงสุด (max heart rate) อาการเจ็บหน้าอกขณะออกกำลังกาย (exercise angina) ระดับการยกตัวของ ST segment เทียบกับ Isoelectric Line (oldpeak) และลักษณะความชันของ ST segment (ST slope)

3.2.2 การกำหนดตัวแปรตาม

ตัวแปรตามในการศึกษานี้คือตัวแปรเชิงคุณภาพที่มี 2 ผลลัพธ์ ได้แก่ เป็นโรคหัวใจ และไม่เป็นโรคหัวใจ

3.3 การเตรียมข้อมูล

กระบวนการเตรียมข้อมูล (Data Cleaning) ก่อนที่จะนำข้อมูลที่ได้จากการหาข้อมูลโดยการสืบค้นบนอินเทอร์เน็ตมาสร้างแบบจำลอง คือการตรวจสอบและจัดการข้อมูลที่หายไป และการแปลงข้อมูลเพื่อให้สามารถนำไปสร้างแบบจำลองได้

3.3.1 กระบวนการตรวจสอบและจัดการข้อมูลที่หายไป

เมื่อตรวจสอบชุดข้อมูล “Heart Disease Dataset” ผู้จัดทำชุดข้อมูลได้จัดการกับข้อมูลที่หายไปโดยการใส่เลข 0 ในกรณีที่ตัวแปรนั้นเป็นข้อมูลเชิงปริมาณหรือตัวแปรนั้นเป็นข้อมูลเชิง

คุณภาพที่ไม่ได้กำหนดให้ผลลัพธ์ 0 มีความหมาย ผู้ศึกษาจึงลบค่าสังเกต (Observation) ดังกล่าวออกทั้งหมด

3.3.2 กระบวนการแปลงข้อมูลเพื่อให้สามารถนำไปสร้างแบบจำลองได้

เนื่องจากตัวแปรเพศ (sex) ลักษณะอาการเจ็บหน้าอก (chest pain type) ระดับน้ำตาลในเลือดในช่วง 2-3 วันที่ผ่านมา โดยงดอาหาร 8 ชั่วโมงก่อนตรวจ (fasting blood sugar) คลื่นไฟฟ้าหัวใจขณะพัก (resting ecg) อาการเจ็บหน้าอกขณะออกกำลังกาย (exercise angina) และลักษณะความชันของ ST segment (ST slope) ที่ได้จากการสืบค้นข้อมูลบนอินเทอร์เน็ตเป็นข้อมูลเชิงคุณภาพจึงต้องแปลงข้อมูลให้เป็นข้อมูลตัวเลขจำนวนเต็ม (Integer Data Type) ก่อนที่จะนำไปสร้างแบบจำลอง เมื่อตรวจสอบชุดข้อมูลพบว่าตัวแปรที่เป็นข้อมูลเชิงคุณภาพทุกตัวแปรได้เป็นข้อมูลตัวเลขจำนวนเต็ม (Integer Data Type) เรียบร้อยแล้ว

3.4 การวิเคราะห์ข้อมูล

3.4.1 ตัวแปรเชิงปริมาณ (Continuous Variable)

คำนวณค่าสถิติเชิงพรรณนา ได้แก่ ค่าเฉลี่ย (Mean) ส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation) ค่าต่ำสุด (Minimum) ค่าควอร์ไทล์ที่ 1 (The First Quartile) ค่าควอร์ไทล์ที่ 2 (The Second Quartile) ค่าควอร์ไทล์ที่ 3 (The Third Quartile) และค่าสูงสุด (Maximum) จากนั้นสร้างกราฟฮีสโทแกรม (Histogram) เพื่อสังเกตลักษณะการกระจายตัวของข้อมูล

3.4.2 ตัวแปรเชิงคุณภาพ (Ordinal หรือ Nominal Scale Variable)

หาค่าสถิติฐานนิยม (Mode) รวมถึงสร้างตารางแจกแจงความถี่ และแผนภูมิวงกลม เพื่ออธิบายลักษณะของกลุ่มตัวอย่าง

3.4.3 เปรียบเทียบลักษณะของข้อมูล เมื่อค่าของตัวแปรเป้าหมาย (Target Class) ต่างกัน

คำนวณค่าเฉลี่ย (Mean) ของตัวแปรคุณลักษณะที่เป็นข้อมูลเชิงปริมาณ และหาฐานนิยมของตัวแปรคุณลักษณะที่เป็นข้อมูลเชิงคุณภาพ เมื่อค่าของตัวแปรเป้าหมาย (Target Class) ต่างกัน

3.5 การสร้างแบบจำลองต้นไม้ตัดสินใจ (Decision Tree Model) พร้อมประเมินประสิทธิภาพ

3.5.1 การหาน้ำหนักความสำคัญของตัวแปร (Feature Importance)

1. กำหนดให้ชุดตัวแปร X คือตัวแปรอิสระ 11 ตัวจากขั้นตอนที่ 3.2.1 และชุดตัวแปร Y คือตัวแปรจากขั้นตอนที่ 3.2.2

2. แบ่งข้อมูลทั้งหมดออกเป็น Training และ Testing Data ด้วยสัดส่วน 70 ต่อ 30

3. สร้างแบบจำลองต้นไม้ตัดสินใจ (Decision Tree) ที่เป็นไปได้ทั้งหมด ซึ่งเกิดจากการจับคู่ระหว่าง criterion ซึ่งจะเป็น Gini หรือ Entropy และ max depth ซึ่งเป็นไปได้ตั้งแต่ 2 ถึง 12 โดยใช้ข้อมูล Training Data และมี random state คือ 1000

4. นำ Testing Data มาประเมินประสิทธิภาพของแบบจำลองทั้ง 22 แบบจำลอง

5. เลือกแบบจำลองที่ดีที่สุด โดยใช้เกณฑ์ Accuracy Score

6. นำแบบจำลองที่ดีที่สุดมาหาน้ำหนักความสำคัญของตัวแปร (Feature Importance)

3.5.2 การหาชุดของตัวแปรและพารามิเตอร์ของแบบจำลองที่ดีที่สุด

1. กำหนดให้ชุดตัวแปร X มีทั้งหมด 10 ชุด โดยคัดตัวแปรที่มีน้ำหนักความสำคัญของตัวแปร (Feature Importance) น้อยที่สุดออก ทำซ้ำจนกระทั้งชุดตัวแปร X สุดท้ายมีตัวแปรคุณลักษณะ 2 ตัว และชุดตัวแปร Y คือตัวแปรจากขั้นตอนที่ 3.2.2

2. แบ่งข้อมูลทั้งหมดออกเป็น Training และ Testing Data ด้วยสัดส่วน 70 ต่อ 30

3. สร้างแบบจำลองต้นไม้ตัดสินใจ (Decision Tree) ซึ่งเกิดจากการจับคู่ระหว่าง criterion ซึ่งจะเป็น Gini หรือ Entropy และ max depth ซึ่งเป็นไปได้ตั้งแต่ 2 ถึง 12 โดยใช้ข้อมูล Training Data และมี random state คือ 1000 ที่เป็นไปได้ทั้งหมดจากชุดตัวแปร X ทั้งหมด 10 ชุด

4. นำ Testing Data มาประเมินประสิทธิภาพของแบบจำลองทั้งหมด

5. เลือกแบบจำลองที่ดีที่สุด โดยใช้เกณฑ์ Accuracy Score

3.5.3 ประเมินประสิทธิภาพของแบบจำลอง

รายงานค่า Precision, Recall, F1-score และ Accuracy ของแบบจำลองที่ดีที่สุดจากขั้นตอนที่ 3.5.2

3.6 การสร้างแบบจำลองป่าไม้สุ่ม (Random Forest Model) พร้อมประเมินประสิทธิภาพ

3.6.1 การหาน้ำหนักความสำคัญของตัวแปร (Feature Importance)

1. กำหนดให้ชุดตัวแปร X คือตัวแปรอิสระ 11 ตัวจากขั้นตอนที่ 3.2.1 และชุดตัวแปร Y คือตัวแปรจากขั้นตอนที่ 3.2.2

2. แบ่งข้อมูลทั้งหมดออกเป็น Training และ Testing Data ด้วยสัดส่วน 70 ต่อ 30

3. สร้างแบบจำลองป่าไม้สุ่ม (Random Forest Model) ที่เป็นไปได้ทั้งหมด ซึ่งมีพารามิเตอร์คือ criterion ซึ่งจะเป็น Gini หรือ Entropy, n estimators ซึ่งเป็นไปได้ตั้งแต่ 5 ถึง 20 และ max depth ซึ่งเป็นไปได้ตั้งแต่ 2 ถึง 12 โดยใช้ข้อมูล Training Data และมี random state คือ 1000

4. นำ Testing Data มาประเมินประสิทธิภาพของแบบจำลองทั้ง 32 แบบจำลอง

5. เลือกแบบจำลองที่ดีที่สุด โดยใช้เกณฑ์ Accuracy Score

6. นำแบบจำลองที่ดีที่สุดมาหาน้ำหนักความสำคัญของตัวแปร (Feature Importance)

3.6.2 การหาชุดของตัวแปรและพารามิเตอร์ของแบบจำลองที่ดีที่สุด

1. กำหนดให้ชุดตัวแปร X มีทั้งหมด 10 ชุด โดยคัดตัวแปรที่มีน้ำหนักความสำคัญของตัวแปร (Feature Importance) น้อยที่สุดออก ทำซ้ำจนกระทั่งชุดตัวแปร X สุดท้ายมีตัวแปรคุณลักษณะ 2 ตัว และชุดตัวแปร Y คือตัวแปรจากขั้นตอนที่ 3.2.2

2. แบ่งข้อมูลทั้งหมดออกเป็น Training และ Testing Data ด้วยสัดส่วน 70 ต่อ 30

3. สร้างแบบจำลองป่าไม้สุ่ม (Random Forest Model) ซึ่งเกิดจากการจับคู่ระหว่าง criterion ซึ่งจะเป็น Gini หรือ Entropy และ n estimators ซึ่งคือจำนวนเลขคี่ตั้งแต่ 5 ถึง 19 โดยใช้ข้อมูล Training Data และมี random state คือ 1000 ที่เป็นไปได้ทั้งหมดจากชุดตัวแปร X ทั้งหมด 10 ชุด

4. นำ Testing Data มาประเมินประสิทธิภาพของแบบจำลองทั้งหมด

5. เลือกแบบจำลองที่ดีที่สุด โดยใช้เกณฑ์ Accuracy Score

3.6.3 ประเมินประสิทธิภาพของแบบจำลอง

รายงานค่า Precision, Recall, F1-score และ Accuracy ของแบบจำลองที่ดีที่สุดจากขั้นตอนที่ 3.6.2

3.7 สรุป และอภิปรายผล

นำแบบจำลองต้นไม้ตัดสินใจ (Decision Tree Model) และแบบจำลองป่าไม้สุ่ม (Random Forest Model) ที่ดีที่สุดมาเปรียบเทียบประสิทธิภาพ โดยใช้เกณฑ์ Precision, Recall, F1-score และ Accuracy เพื่อเลือกแบบจำลองที่ดีที่สุด

บทที่ 4

ผลการดำเนินการวิจัย

4.1 ข้อมูลที่ได้จากการรวบรวมเพื่อใช้สร้างแบบจำลอง

รูปที่ 4.1 ตัวอย่างข้อมูลที่ได้จากการสืบค้นบนอินเทอร์เน็ตและนำไปสร้างแบบจำลอง

	age	sex	chest pain type	resting bp s	cholesterol	fasting blood sugar	resting ecg	max heart rate	exercise angina	oldpeak	ST slope	target
0	40	1	2	140	289	0	0	172	0	0.0	1	0
1	49	0	3	160	180	0	0	156	0	1.0	2	1
2	37	1	2	130	283	0	1	98	0	0.0	1	0
3	48	0	4	138	214	0	0	108	1	1.5	2	1
4	54	1	3	150	195	0	0	122	0	0.0	1	0
5	39	1	3	120	339	0	0	170	0	0.0	1	0
6	45	0	2	130	237	0	0	170	0	0.0	1	0
7	54	1	2	110	208	0	0	142	0	0.0	1	0
8	37	1	4	140	207	0	0	130	1	1.5	2	1
9	48	0	2	120	284	0	0	120	0	0.0	1	0

จากการหาข้อมูลโดยการสืบค้นบนอินเทอร์เน็ต คือ ชุดข้อมูล “Heart Disease Dataset” ซึ่งประกอบด้วยข้อมูลจำนวน 1190 ตัวอย่าง

4.2 ตัวแปรที่ได้จากการกำหนด

ตารางที่ 4.1 ตัวแปรคุณลักษณะ ตัวแปรเป้าหมาย และเงื่อนไขที่ใช้ในการสร้างต้นไม้ตัดสินใจ

No.	Attributes	Type	Description
1	age	Continuous	อายุ (ปี)
2	sex	Nominal	เพศ {0: หญิง, 1: ชาย}
3	chest pain type	Ordinal	อาการเจ็บหน้าอกแบบ Typical angina มี 3 ลักษณะ ได้แก่ 1) เจ็บแน่นใต้หน้าอกและอาจร้าวไปกราม แขน คอ หรือไหล่ 2) ถูกกระตุ้นโดยการออกแรงหรือภาวะเครียด 3) อาการดีขึ้นเมื่อพักหรือได้ยา Nitroglycerine {1: มีครบ 3 ลักษณะ, 2: มี 2 ลักษณะ, 3: มี 0-1 ลักษณะ (0 ลักษณะคือมีอาการเจ็บหน้าอกแต่ไม่อยู่ใน 3 ลักษณะข้างต้น), 4: ไม่อาการเจ็บหน้าอก}
4	resting bp s	Continuous	ความดันโลหิตขณะพัก (มิลลิเมตรปรอท)
5	cholesterol	Continuous	ระดับของคอเลสเตอรอลในเลือด (มิลลิกรัมต่อเดซิลิตร)

No.	Attributes	Type	Description
6	fasting blood sugar	Ordinal	ระดับน้ำตาลในเลือดในช่วง 2-3 วันที่ผ่านมา โดยงดอาหาร 8 ชั่วโมงก่อนตรวจ {0: ค่าน้ำตาลกลูโคสต่ำกว่าหรือเท่ากับ 120 มิลลิกรัมต่อเดซิลิตร, 1: ค่าน้ำตาลกลูโคสต่ำกว่าสูงกว่า 120 มิลลิกรัมต่อเดซิลิตร}
7	resting ecg	Nominal	คลื่นไฟฟ้าหัวใจขณะพัก {0: ปกติ, 1: ST-T wave ผิดปกติ, 2: มีโอกาสเป็นภาวะหัวใจห้องล่างซ้ายหนา}
8	max heart rate	Continuous	อัตราการเต้นของหัวใจสูงสุด (ครั้งต่อนาที)
9	exercise angina	Nominal	มีอาการเจ็บหน้าอกขณะออกกำลังกาย {0: ไม่มี, 1: มี}
10	oldpeak	Continuous	ระดับการยกตัวของ ST segment เทียบกับ Isoelectric Line (มิลลิเมตร)
11	ST slope	Nominal	ลักษณะความชันของ ST segment {1: ความชันเป็นบวก, 2: ความชันเป็นศูนย์, 3: ความชันเป็นลบ}
12	**target (Target Class)**	**Nominal**	**เป็นโรคหัวใจ {0: ไม่ใช่, 1: ใช่}**

4.3 ผลการเตรียมข้อมูล

ตารางที่ 4.2 Data Type และ Missing Value ของข้อมูลก่อนและหลังการจัดการ

Variables	DType	Number of Original Missing Values	Number of Missing Values After Drop Na
age	int64	0	0
sex	int64	0	0
chest pain type	int64	0	0
resting bp s	int64	1	0
cholesterol	int64	172	0
fasting blood sugar	int64	0	0
resting ecg	int64	0	0
max heart rate	int64	0	0
exercise angina	int64	0	0

Variables	DType	Number of Original Missing Values	Number of Missing Values After Drop Na
oldpeak	float64	0	0
ST slope	int64	1	0
target	int64	0	0

4.4 ผลการวิเคราะห์ข้อมูล

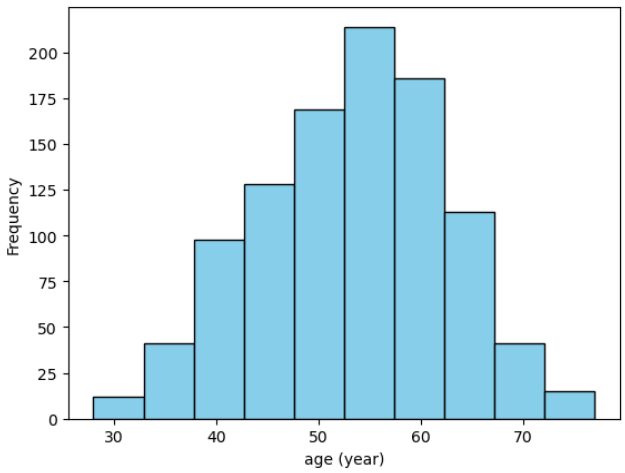
4.4.1 ตัวแปรเชิงปริมาณ (Continuous Variable)

ตารางที่ 4.3 สถิติเชิงพรรณนา (Descriptive Statistics) ของตัวแปรเชิงปริมาณ

	Age (อายุ: ปี)	resting bp s (ความดันโลหิตขณะพัก : มิลลิเมตรปรอท)	cholesterol (ระดับของคอเลสเตอรอลในเลือด: มิลลิกรัมต่อเดซิลิตร)
Mean	53.28	132.55	245.96
Standard Deviation	9.41	17.45	57.25
Minimum	28	92	85
The First Quartile	46	120	209
The Second Quartile	54	130	240
The Third Quartile	60	140	276
Maximum	77	200	603

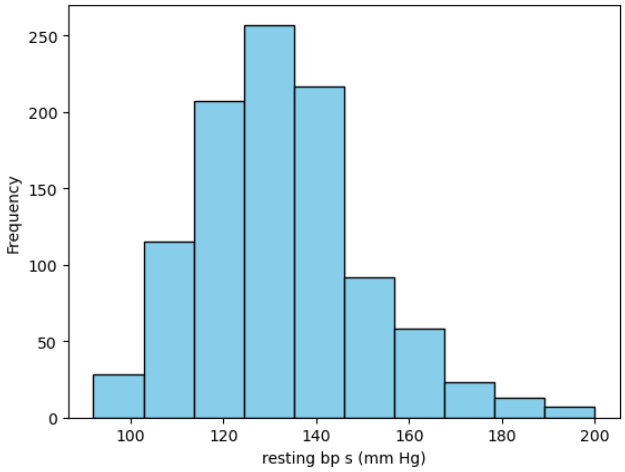
	max heart rate (อัตราการเต้นของหัวใจสูงสุด: ครั้งต่อนาที)	oldpeak (ระดับการยกตัวของ ST segment เทียบกับ Isoelectric Line: มิลลิเมตร)
Mean	142.74	0.94
Standard Deviation	24.52	1.09
Minimum	69	-0.1
The First Quartile	125	0
The Second Quartile	144	0.6
The Third Quartile	161	1.6
Maximum	202	6.2

รูปที่ 4.2 ลักษณะการกระจายของตัวแปร age (อายุ)



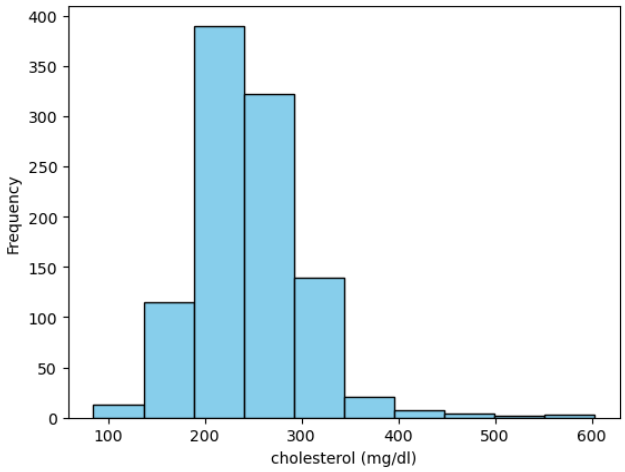
จากตารางที่ 4.3 และรูปที่ 4.2 จะเห็นได้ว่าข้อมูลอายุ (age) มีการกระจายตัวของข้อมูลในลักษณะโค้งปกติ โดยกลุ่มตัวอย่างมีอายุเฉลี่ยที่ประมาณ 53.28 ปี ซึ่งใกล้เคียงกับค่ามัธยฐานที่ 54 ปี อายุที่น้อยที่สุดของตัวอย่างคือ 28 ปี อายุที่สูงที่สุดของตัวอย่างคือ 77 ปี และมีส่วนเบี่ยงเบนมาตรฐานที่ 9.41 โดยตัวอย่างมีการกระจายหนาแน่นในช่วงอายุ 50 ถึง 60 ปี

รูปที่ 4.3 ลักษณะการกระจายของตัวแปร resting bp s (ความดันโลหิตขณะพัก)



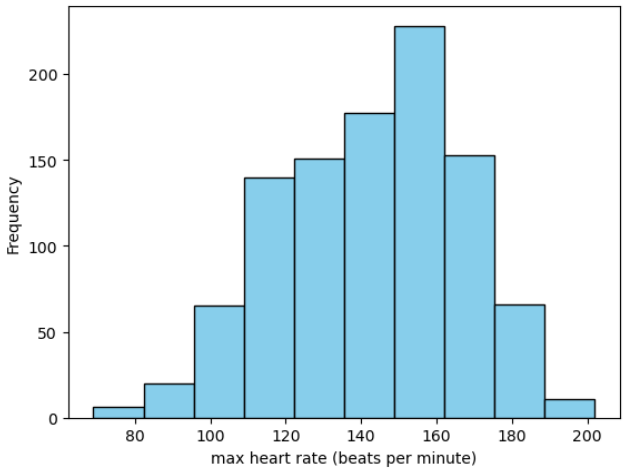
จากตารางที่ 4.3 และรูปที่ 4.3 จะเห็นได้ว่าข้อมูลความดันโลหิตขณะพัก (resting bp s) มีการกระจายตัวของข้อมูลในลักษณะเบ้ขวา โดยกลุ่มตัวอย่างมีความดันโลหิตขณะพักเฉลี่ยที่ประมาณ 132.55 มิลลิเมตรปรอท ซึ่งใกล้เคียงกับค่ามัธยฐานที่ 130 มิลลิเมตรปรอท ความดันโลหิตขณะพักที่น้อยที่สุดของตัวอย่างคือ 92 มิลลิเมตรปรอท ความดันโลหิตขณะพักที่สูงที่สุดของตัวอย่างคือ 200 มิลลิเมตรปรอท และมีส่วนเบี่ยงเบนมาตรฐานที่ 17.45 โดยตัวอย่างมีการกระจายหนาแน่นในช่วง 120 ถึง 140 มิลลิเมตรปรอท

รูปที่ 4.4 ลักษณะการกระจายของตัวแปร cholesterol (ระดับของคอเลสเตอรอลในเลือด)



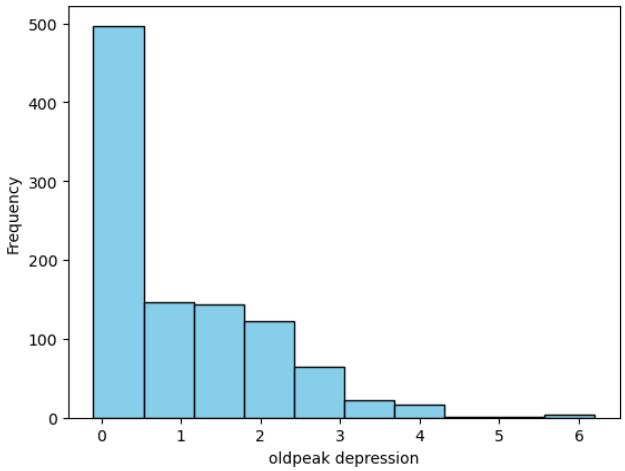
จากตารางที่ 4.3 และรูปที่ 4.4 จะเห็นได้ว่าข้อมูลระดับของคอเลสเตอรอลในเลือด (cholesterol) มีการกระจายตัวของข้อมูลในลักษณะเบ้ขวา โดยกลุ่มตัวอย่างมีระดับของคอเลสเตอรอลในเลือดเฉลี่ยที่ประมาณ 245.96 มิลลิกรัมต่อเดซิลิตร ซึ่งใกล้เคียงกับค่ามัธยฐานที่ 240 มิลลิกรัมต่อเดซิลิตร ระดับของคอเลสเตอรอลในเลือดที่น้อยที่สุดของตัวอย่างคือ 85 มิลลิกรัมต่อเดซิลิตร ระดับของคอเลสเตอรอลในเลือดที่สูงที่สุดของตัวอย่างคือ 603 มิลลิกรัมต่อเดซิลิตร และมีส่วนเบี่ยงเบนมาตรฐานที่ 57.25 โดยตัวอย่างมีการกระจายหนาแน่นในช่วง 200 ถึง 300 มิลลิกรัมต่อเดซิลิตร

รูปที่ 4.5 ลักษณะการกระจายของตัวแปร max heart rate (อัตราการเต้นของหัวใจสูงสุด)



จากตารางที่ 4.3 และรูปที่ 4.5 จะเห็นได้ว่าข้อมูลอัตราการเต้นของหัวใจสูงสุด (max heart rate) มีการกระจายตัวของข้อมูลในลักษณะเบ้ซ้าย โดยกลุ่มตัวอย่างมีอัตราการเต้นของหัวใจสูงสุดเฉลี่ยที่ประมาณ 142.74 ครั้งต่อนาที ซึ่งใกล้เคียงกับค่ามัธยฐานที่ 144 ครั้งต่อนาที อัตราการเต้นของหัวใจสูงสุดที่น้อยที่สุดของตัวอย่างคือ 69 ครั้งต่อนาที อัตราการเต้นของหัวใจสูงสุดที่สูงที่สุดของตัวอย่างคือ 202 ครั้งต่อนาที และมีส่วนเบี่ยงเบนมาตรฐานที่ 24.52 โดยตัวอย่างมีการกระจายหนาแน่นในช่วง 140 ถึง 160 ครั้งต่อนาที

รูปที่ 4.6 ลักษณะการกระจายของตัวแปร oldpeak (ระดับการยกตัวของ ST segment เทียบกับ Isoelectric Line)

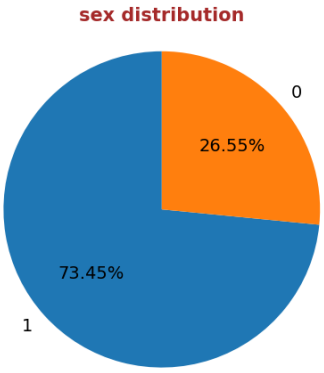


จากตารางที่ 4.3 และรูปที่ 4.6 จะเห็นได้ว่าข้อมูลระดับการยกตัวของ ST segment เทียบกับ Isoelectric Line (oldpeak) มีการกระจายตัวของข้อมูลในลักษณะเบ้ขวา โดยกลุ่มตัวอย่างมีระดับการยกตัวของ ST segment เทียบกับ Isoelectric Line เฉลี่ยที่ประมาณ 0.94 มิลลิเมตร ระดับการยกตัวของ ST segment เทียบกับ Isoelectric Line ที่น้อยที่สุดของตัวอย่างคือ -0.1 มิลลิเมตร ระดับการยกตัวของ ST segment เทียบกับ Isoelectric Line ที่สูงที่สุดของตัวอย่างคือ 6.2 มิลลิเมตร และมีส่วนเบี่ยงเบนมาตรฐานที่ 1.09 โดยตัวอย่างมีการกระจายหนาแน่นในช่วง 0 มิลลิเมตร

4.4.2 ตัวแปรเชิงคุณภาพ (Ordinal หรือ Nominal Scale Variable)

ตารางที่ 4.4 ตารางแจกแจงความถี่ของตัวแปร sex และรูปที่ 4.7 แผนภูมิวงกลมแสดงสัดส่วนความถี่ของตัวแปร sex

sex	
Labels	Frequency
0	270
1	747

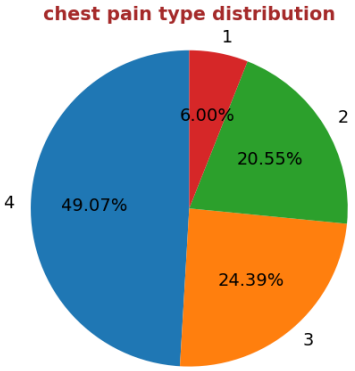


ตัวแปร sex (เพศ) มีผลลัพธ์ 2 ค่า คือ 0: หญิง และ 1:ชาย โดยตัวอย่างส่วนใหญ่ของข้อมูลชุดนี้เป็นเพศชาย ซึ่งมีจำนวน 747 คน คิดเป็นร้อยละ 73.45 ของกลุ่มตัวอย่างทั้งหมด ขณะที่เพศหญิงมีจำนวน 270 คน คิดเป็นร้อยละ 26.55 ของกลุ่มตัวอย่างทั้งหมด

ตารางที่ 4.5 ตารางแจกแจงความถี่ของตัวแปร chest pain type และ

รูปที่ 4.8 แผนภูมิวงกลมแสดงสัดส่วนความถี่ของตัวแปร chest pain type

chest pain type	
Labels	Frequency
1	61
2	209
3	248
4	499

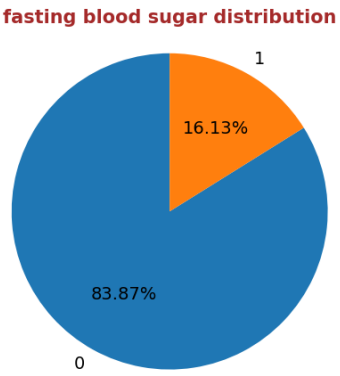


เนื่องจากอาการเจ็บหน้าอกแบบ Typical angina มี 3 ลักษณะ ได้แก่ 1) เจ็บแน่นใต้หน้าอกและอาจร้าวไปกราม แขน คอ หรือไหล่ 2) ถูกกระตุ้นโดยการออกแรงหรือภาวะเครียด 3) อาการดีขึ้นเมื่อพักหรือได้ยา Nitroglycerine ตัวแปร chest pain type (อาการเจ็บหน้าอก) จึงมีผลลัพธ์ 4 ค่า คือ 1: มีครบ 3 ลักษณะของอาการเจ็บหน้าอกแบบ Typical angina, 2: มี 2 ลักษณะอาการเจ็บหน้าอกแบบ Typical angina, 3: มี 0-1 ลักษณะอาการเจ็บหน้าอกแบบ Typical angina (0 ลักษณะอาการเจ็บหน้าอกแบบ Typical angina คือ มีอาการเจ็บหน้าอกแต่ไม่อยู่ใน 3 ลักษณะข้างต้น) และ 4: ไม่อาการเจ็บหน้าอก โดยตัวอย่างส่วนใหญ่ของข้อมูลชุดนี้ไม่มีอาการเจ็บหน้าอก ซึ่งมีจำนวน 499 คน คิดเป็นร้อยละ 49.07 ของกลุ่มตัวอย่างทั้งหมด ขณะที่มีตัวอย่างที่มีครบ 3 ลักษณะของอาการเจ็บหน้าอกแบบ Typical angina จำนวน 61 คน คิดเป็นร้อยละ 6.00 ของกลุ่มตัวอย่างทั้งหมด ตัวอย่างที่มี 2 ลักษณะของอาการเจ็บหน้าอกแบบ Typical angina จำนวน 209 คน คิดเป็นร้อยละ 20.55 ของกลุ่มตัวอย่างทั้งหมด และตัวอย่างที่มี 0-1 ลักษณะของอาการเจ็บหน้าอกแบบ Typical angina จำนวน 248 คน คิดเป็นร้อยละ 24.39 ของกลุ่มตัวอย่างทั้งหมด

ตารางที่ 4.6 ตารางแจกแจงความถี่ของตัวแปร fasting blood sugar และ

รูปที่ 4.9 แผนภูมิวงกลมแสดงสัดส่วนความถี่ของตัวแปร fasting blood sugar

fasting blood sugar	
Labels	Frequency
0	853
1	164

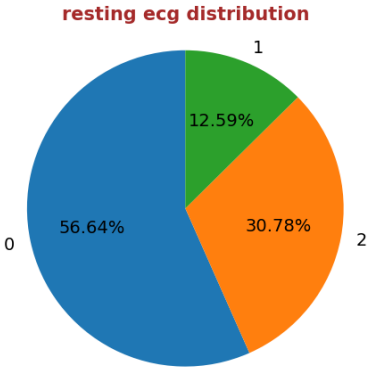


ตัวแปร fasting blood sugar (ระดับน้ำตาลในเลือดในช่วง 2-3 วันที่ผ่านมา โดยงดอาหาร 8 ชั่วโมงก่อนตรวจ) มีผลลัพธ์ 2 ค่า คือ 0: ค่าน้ำตาลกลูโคสต่ำกว่าหรือเท่ากับ 120 มิลลิกรัมต่อเดซิลิตร และ 1: ค่าน้ำตาลกลูโคสต่ำกว่าสูงกว่า 120 มิลลิกรัมต่อเดซิลิตร โดยตัวอย่างส่วนใหญ่ของข้อมูลชุดนี้มีค่าน้ำตาลกลูโคสต่ำกว่าหรือเท่ากับ 120 มิลลิกรัมต่อเดซิลิตร ซึ่งมีจำนวน 853 คน คิดเป็นร้อยละ 83.87 ของกลุ่มตัวอย่างทั้งหมด ขณะที่ตัวอย่างที่มีค่าน้ำตาลกลูโคสต่ำกว่าสูงกว่า 120 มิลลิกรัมต่อเดซิลิตรมีจำนวน 164 คน คิดเป็นร้อยละ 16.13 ของกลุ่มตัวอย่างทั้งหมด

ตารางที่ 4.7 ตารางแจกแจงความถี่ของตัวแปร resting ecg และ

รูปที่ 4.10 แผนภูมิวงกลมแสดงสัดส่วนความถี่ของตัวแปร resting ecg

resting ecg	
Labels	Frequency
0	576
1	128
2	313

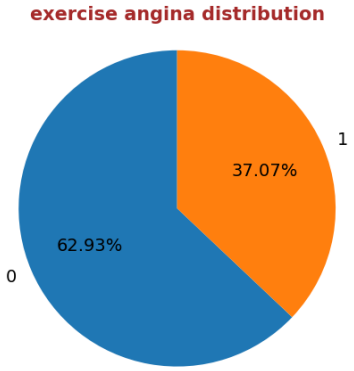


ตัวแปร resting ecg (คลื่นไฟฟ้าหัวใจขณะพัก) มีผลลัพธ์ 3 ค่า คือ 0: ปกติ, 1: ST-T wave ผิดปกติ และ 2: มีโอกาสเป็นภาวะหัวใจห้องล่างซ้ายหนา โดยตัวอย่างส่วนใหญ่ของข้อมูลชุดนี้มีคลื่นไฟฟ้าหัวใจขณะพักปกติ ซึ่งมีจำนวน 576 คน คิดเป็นร้อยละ 56.64 ของกลุ่มตัวอย่างทั้งหมด ขณะที่ตัวอย่างที่มี ST-T wave ผิดปกติมีจำนวน 128 คน คิดเป็นร้อยละ 12.59 ของกลุ่มตัวอย่างทั้งหมด และตัวอย่างที่มีโอกาสเป็นภาวะหัวใจห้องล่างซ้ายหนามีจำนวน 313 คน คิดเป็นร้อยละ 30.78 ของกลุ่มตัวอย่างทั้งหมด

ตารางที่ 4.8 ตารางแจกแจงความถี่ของตัวแปร exercise angina และ

รูปที่ 4.11 แผนภูมิวงกลมแสดงสัดส่วนความถี่ของตัวแปร exercise angina

exercise angina	
Labels	Frequency
0	640
1	377

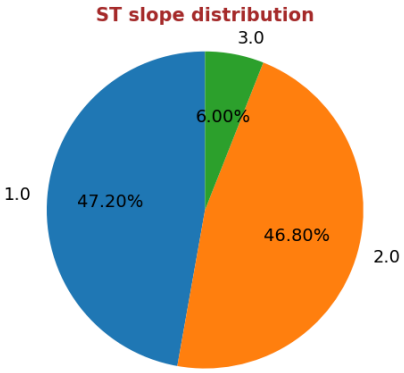


ตัวแปร exercise angina (อาการเจ็บปวดขณะออกกำลังกาย) มีผลลัพธ์ 2 ค่า คือ 0: ไม่มีอาการเจ็บปวดขณะออกกำลังกาย และ 1: มีอาการเจ็บปวดขณะออกกำลังกาย โดยตัวอย่างส่วนใหญ่ของข้อมูลชุดนี้ไม่มีอาการเจ็บปวดขณะออกกำลังกาย ซึ่งมีจำนวน 640 คน คิดเป็นร้อยละ 62.93 ของกลุ่มตัวอย่างทั้งหมด ขณะที่ตัวอย่างที่มีอาการเจ็บปวดขณะออกกำลังกาย มีจำนวน 377 คน คิดเป็นร้อยละ 37.07 ของกลุ่มตัวอย่างทั้งหมด

ตารางที่ 4.9 ตารางแจกแจงความถี่ของตัวแปร ST slope และ

รูปที่ 4.12 แผนภูมิวงกลมแสดงสัดส่วนความถี่ของตัวแปร ST slope

ST slope	
Labels	Frequency
1	480
2	476
3	61

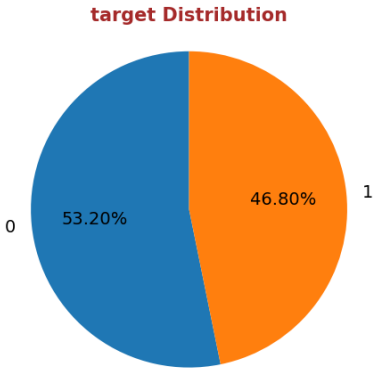


ตัวแปร ST slope (ลักษณะความชันของ ST segment) มีผลลัพธ์ 3 ค่า คือ 1: ST segment มีความชันเป็นบวก, 2: ST segment มีความชันเป็นศูนย์ และ 3: ST segment มีความชันเป็นลบ โดยตัวอย่างส่วนใหญ่ของข้อมูลชุดนี้ ST segment มีความชันเป็นบวก ซึ่งมีจำนวน 480 คน คิดเป็นร้อยละ 47.20 ของกลุ่มตัวอย่างทั้งหมด ขณะที่ตัวอย่างที่ ST segment มีความชันเป็นศูนย์มีจำนวน 476 คน คิดเป็นร้อยละ 46.80 ของกลุ่มตัวอย่างทั้งหมด และตัวอย่างที่มี ST segment มีความชันเป็นลบมีจำนวน 61 คน คิดเป็นร้อยละ 6.00 ของกลุ่มตัวอย่างทั้งหมด

ตารางที่ 4.10 ตารางแจกแจงความถี่ของตัวแปร target และ

รูปที่ 4.13 แผนภูมิวงกลมแสดงสัดส่วนความถี่ของตัวแปร target

target	
Labels	Frequency
0	541
1	476



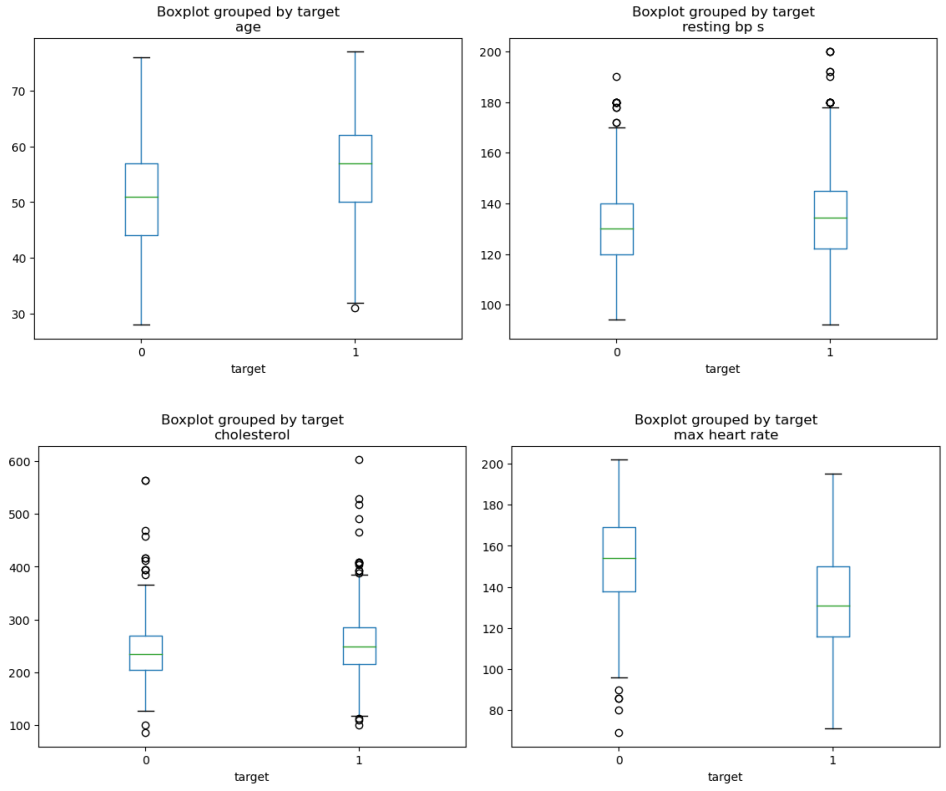
ตัวแปร target (โรคหัวใจ) มีผลลัพธ์ 2 ค่า คือ 0: ไม่เป็นโรคหัวใจ และ 1: เป็นโรคหัวใจ โดยตัวอย่างส่วนใหญ่ของข้อมูลชุดนี้ไม่เป็นโรคหัวใจ ซึ่งมีจำนวน 541 คน คิดเป็นร้อยละ 53.20 ของกลุ่มตัวอย่างทั้งหมด ขณะที่ตัวอย่างที่เป็นโรคหัวใจ มีจำนวน 476 คน คิดเป็นร้อยละ 46.80 ของกลุ่มตัวอย่างทั้งหมด

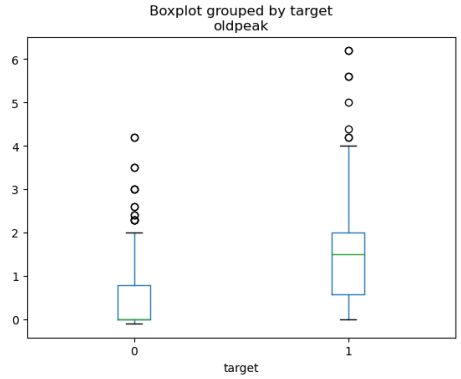
4.4.3 เปรียบเทียบลักษณะของข้อมูล เมื่อค่าของตัวแปรเป้าหมาย (Target Class) ต่างกัน

ตารางที่ 4.11 ค่าเฉลี่ยของตัวแปรเชิงปริมาณ เมื่อ target เป็น 0 และ 1

target	age (year)	resting bp s (mm Hg)	cholesterol (mg/dl)	max heart rate (beats per minute)	oldpeak (mm)
0: ไม่เป็นโรคหัวใจ	50.87	129.77	240.22	151.65	0.46
1: เป็นโรคหัวใจ	56.02	135.72	252.47	132.62	1.49

รูปที่ 4.14 Box plot ของตัวแปรเชิงปริมาณ เมื่อ target เป็น 0 และ 1





ในด้านอายุ (age) ความดันโลหิตขณะพัก (resting bp s) ระดับของคอเลสเตอรอลในเลือด (cholesterol) และระดับการยกตัวของ ST segment เทียบกับ Isoelectric Line (oldpeak) ตัวอย่างที่เป็นโรคหัวใจจะเกาะกลุ่มอยู่ในระดับที่สูงกว่า ตัวอย่างที่ไม่เป็นโรคหัวใจ แต่ในด้านอัตราการเต้นของหัวใจสูงสุด (max heart rate) ตัวอย่างที่เป็นโรคหัวใจจะเกาะกลุ่มอยู่ในระดับที่ต่ำกว่า ตัวอย่างที่ไม่เป็นโรคหัวใจ

ตารางที่ 4.12 ฐานนิยมของตัวแปรเชิงคุณภาพ เมื่อ Target เป็น 0 และ 1

target	**0: ไม่เป็นโรคหัวใจ**	**1: เป็นโรคหัวใจ**
sex	1	1
chest pain type	3	4
fasting blood sugar	0	0
resting ecg	0	0
exercise angina	0	1
ST slope	1	2

จากตารางที่ 4.12 จะเห็นได้ว่าฐานนิยมของตัวแปรเชิงคุณภาพที่ไม่แตกต่างกัน ได้แก่ เป็นเพศชาย มีค่าน้ำตาลกลูโคสต่ำกว่าหรือเท่ากับ 120 มิลลิกรัมต่อเดซิลิตร มีคลื่นไฟฟ้าหัวใจขณะพักปกติ ขณะที่ฐานนิยมของตัวแปรเชิงคุณภาพที่แตกต่างกัน กลุ่มตัวอย่างที่ไม่เป็นโรคหัวใจมีฐานนิยมคือมี 0-1 ลักษณะของอาการเจ็บหน้าอกแบบ Typical angina ไม่มีอาการเจ็บหน้าอกขณะออกกำลังกาย และความชันของ ST segment เป็นบวก แต่กลุ่มตัวอย่างที่เป็นโรคหัวใจมีฐานนิยมคือมีอาการเจ็บหน้าอกขณะออกกำลังกาย และความชันของ ST segment เท่ากับศูนย์

4.5 แบบจำลองต้นไม้ตัดสินใจ (Decision Tree Model) และประสิทธิภาพของแบบจำลอง

4.5.1 น้ำหนักความสำคัญของตัวแปร (Feature Importance)

แบบจำลองต้นไม้ตัดสินใจที่ดีที่สุด จากแบบจำลองทั้งหมด 22 แบบ ซึ่งเกิดจากการจับคู่ระหว่าง criterion ซึ่งจะเป็น Gini หรือ Entropy และ max depth ซึ่งเป็นไปได้ตั้งแต่ 2 ถึง 12

สำหรับชุดตัวแปรที่ประกอบด้วยตัวแปรคุณลักษณะ (Feature) ทั้งหมด 11 ตัว ด้วยเกณฑ์ Accuracy Score คือ

DecisionTreeClassifier ที่มีพารามิเตอร์

1. criterion = ‘gini’
2. max_depth = 12
3. random_state = 1000

ด้วยคะแนน Accuracy Score หรือคะแนนความแม่นยำสูงสุดที่ 90.52%

หลังทำการแบ่งข้อมูลเป็น Training และ Testing Data ด้วยสัดส่วน 70 ต่อ 30 จากนั้นจัดลำดับน้ำหนักความสำคัญของตัวแปรคุณลักษณะแต่ละตัว (Feature Importance) ได้ผลดังนี้

ตารางที่ 4.13 น้ำหนักความสำคัญของตัวแปรคุณลักษณะแต่ละตัว (Feature Importance) เรียงจากมากไปน้อย

Features	**Feature Importance**
ST slope	0.360424
chest pain type	0.132422
cholesterol	0.083976
max heart rate	0.074966
age	0.073027
oldpeak	0.072709
resting bp s	0.072053
sex	0.040893
fasting blood sugar	0.031367
exercise angina	0.030427
resting ecg	0.027737
Total	**1.000000**

4.5.2 ชุดของตัวแปรและพารามิเตอร์ของแบบจำลองที่ดีที่สุด

4.5.2.1 ชุดของตัวแปร

ชุดของตัวแปรทั้ง 10 ชุด โดยจะคัดตัวแปรที่มีน้ำหนักความสำคัญของตัวแปร (Feature Importance) น้อยที่สุดออก ได้แก่

X_1 = ['age', 'sex', 'chest pain type', 'resting bp s', 'cholesterol', 'fasting blood sugar', 'resting ecg', 'max heart rate', 'exercise angina', 'oldpeak', 'ST slope']

X_2 = ['age', 'sex', 'chest pain type', 'resting bp s', 'cholesterol', 'fasting blood sugar', 'max heart rate', 'exercise angina', 'oldpeak', 'ST slope']

X_3 = ['age', 'sex', 'chest pain type', 'resting bp s', 'cholesterol', 'fasting blood sugar', 'max heart rate', 'oldpeak', 'ST slope']

X_4 = ['age', 'sex', 'chest pain type', 'resting bp s', 'cholesterol', 'max heart rate', 'oldpeak', 'ST slope']

X_5 = ['age', 'chest pain type', 'resting bp s', 'cholesterol', 'max heart rate', 'oldpeak', 'ST slope']

X_6 = ['age', 'chest pain type', 'cholesterol', 'max heart rate', 'oldpeak', 'ST slope']

X_7 = ['age', 'chest pain type', 'cholesterol', 'max heart rate', 'ST slope']

X_8 = ['chest pain type', 'cholesterol', 'max heart rate', 'ST slope']

X_9 = ['chest pain type', 'cholesterol', 'ST slope']

X_10 = ['chest pain type', 'ST slope']

4.5.2.2 แบบจำลองที่ดีที่สุดจากแต่ละชุดของตัวแปร

ตารางที่ 4.14 แบบจำลองที่ดีที่สุดจากแต่ละชุดของตัวแปร

ชุดของตัวแปร	Model (random_state = 1000)	Accuracy Score
X_1	criterion= 'gini', max_depth=12	0.905229
X_2	criterion= 'gini', max_depth=12	0.905229
X_3	criterion= 'gini', max_depth=12	0.888889
X_4	criterion='entropy', max_depth=12	0.882353
X_5	criterion='entropy', max_depth=10	0.885621
X_6	criterion= 'gini', max_depth=10	0.875817

ชุดของตัวแปร	Model (random_state = 1000)	Accuracy Score
X_7	criterion= 'gini', max_depth=10	0.859477
X_8	criterion='entropy', max_depth=12	0.843137
X_9	criterion='entropy', max_depth=7	0.833333
X_10	criterion='entropy', max_depth=3	0.787582

จากตารางที่ 4.14 จะเห็นได้ว่า ชุดของตัวแปรที่ดีที่สุดของแบบจำลองที่ดีที่สุดคือ ชุดของข้อมูลที่ X_2 ซึ่งประกอบด้วยตัวแปร age, sex, chest pain type, resting bp s, cholesterol, fasting blood sugar, max heart rate, exercise angina, oldpeak และ ST slope โดยมี Accuracy Score เท่ากับ 0.905229 ซึ่งเป็น DecisionTreeClassifier ที่มีพารามิเตอร์

1. criterion = ‘Gini’
2. max_depth = 12
3. random_state = 1000

4.5.3 ประสิทธิภาพของแบบจำลอง

ตารางที่ 4.15 ประสิทธิภาพของแบบจำลอง ทดสอบด้วย Training Data

test	Precision	Recall	F1-score	Support
Accuracy	0.9986			711
0	0.9975	1	0.9988	402
1	1	0.9968	0.9984	309
Macro avg	0.9988	0.9984	0.9986	711
Weighted avg	0.9986	0.9986	0.9986	711

Accuracy = 0.9986 หมายความว่า จำนวนการคาดการณ์ที่ถูกต้องทุกประเภท คิดเป็นร้อยละ 99.86 ของจำนวนการคาดการณ์ทั้งหมด

Precision

[Target = 0] = 0.9975 หมายความว่า ในจำนวนคนที่แบบจำลองคาดการณ์ว่าไม่เป็นโรคหัวใจทั้งหมด มีคนที่ไม่ได้เป็นโรคหัวใจจริงร้อยละ 99.75

[Target = 1] = 1 หมายความว่า ในจำนวนคนที่แบบจำลองคาดการณ์ว่าเป็นโรคหัวใจทั้งหมด มีคนที่เป็นโรคหัวใจจริงร้อยละ 100

Recall

[Target = 0] = 1 หมายความว่า ในจำนวนคนที่ไม่เป็นโรคหัวใจจริงทั้งหมด แบบจำลองคาดการณ์ว่าไม่เป็นโรคหัวใจร้อยละ 100

[Target = 1] = 0.9968 หมายความว่า ในจำนวนคนที่เป็นโรคหัวใจจริงทั้งหมด แบบจำลองคาดการณ์ว่าเป็นโรคหัวใจร้อยละ 99.68

F1-score

[Target = 0] = 0.9988 หมายความว่า เมื่อพิจารณาคนที่ไม่เป็นโรคหัวใจจริง ค่าเฉลี่ยฮาร์โมนิคของ Precision และ Recall คือร้อยละ 99.88

[Target = 1] = 0.9986 หมายความว่า เมื่อพิจารณาคนที่เป็นโรคหัวใจจริง ค่าเฉลี่ยฮาร์โมนิคของ Precision และ Recall คือร้อยละ 99.86

ตารางที่ 4.16 ประสิทธิภาพของแบบจำลอง ทดสอบด้วย Testing Data

Train	Precision	Recall	F1-score	Support
Accuracy		0.9052		306
0	0.8929	0.8993	0.8961	139
1	0.9157	0.9192	0.9129	167
Macro avg	0.9043	0.9047	0.9045	306
Weighted avg	0.9053	0.9052	0.9053	306

Accuracy = 0.9052 หมายความว่า จำนวนการคาดการณ์ที่ถูกต้องทุกประเภท คิดเป็นร้อยละ 90.52 ของจำนวนการคาดการณ์ทั้งหมด

Precision

[Target = 0] = 0.8929 หมายความว่า ในจำนวนคนที่แบบจำลองคาดการณ์ว่าไม่เป็นโรคหัวใจทั้งหมด มีคนที่ไม่ได้เป็นโรคหัวใจจริงร้อยละ 89.29

[Target = 1] = 0.9157 หมายความว่า ในจำนวนคนที่แบบจำลองคาดการณ์ว่าเป็นโรคหัวใจทั้งหมด มีคนที่เป็นโรคหัวใจจริงร้อยละ 91.57

Recall

[Target = 0] = 0.8993 หมายความว่า ในจำนวนคนที่ไม่เป็นโรคหัวใจจริงทั้งหมด แบบจำลองคาดการณ์ว่าไม่เป็นโรคหัวใจร้อยละ 89.93

[Target = 1] = 0.9102 หมายความว่า ในจำนวนคนที่เป็นโรคหัวใจจริงทั้งหมด แบบจำลองคาดการณ์ว่าเป็นโรคหัวใจร้อยละ 91.02

F1-score

[Target = 0] = 0.8961 หมายความว่า เมื่อพิจารณาคนที่ไม่เป็นโรคหัวใจจริง ค่าเฉลี่ยฮาร์โมนิคของ Precision และ Recall คือร้อยละ 89.61

[Target = 1] = 0.9129 หมายความว่า เมื่อพิจารณาคนที่เป็นโรคหัวใจจริง ค่าเฉลี่ยฮาร์โมนิคของ Precision และ Recall คือร้อยละ 91.29

4.6 แบบจำลองป่าไม้สุ่ม (Random Forest)

4.6.1 น้ำหนักความสำคัญของตัวแปรแต่ละชนิด (Feature Importance)

แบบจำลองป่าไม้สุ่มที่ดีที่สุด จากแบบจำลองทั้งหมด 176 แบบจำลอง ซึ่งมีพารามิเตอร์คือ criterion ซึ่งจะเป็น Gini หรือ Entropy, n estimators ซึ่งคือจำนวนเลขคี่ตั้งแต่ 5 ถึง 19 และ max depth ซึ่งเป็นไปได้ตั้งแต่ 2 ถึง 12 สำหรับชุดตัวแปรที่ประกอบด้วยตัวแปรคุณลักษณะ (Feature) ทั้งหมด 11 ตัว ด้วยเกณฑ์ Accuracy Score คือ

RandomForestClassifier ที่มีพารามิเตอร์

1. criterion = ‘gini’
2. n_estimators = 17
3. max_depth = 12
4. random_state = 1000

ด้วยคะแนน Accuracy Score หรือคะแนนความแม่นยำสูงสุดที่ 90.85%

หลังจากการแบ่งข้อมูลให้เป็น Training และ Testing Data ด้วยอัตราส่วน 70 ต่อ 30 จากนั้นจัดลำดับน้ำหนักความสำคัญของตัวแปรแต่ละตัว (Feature Importance)ได้ผลดังนี้

ตารางที่ 4.17 น้ำหนักความสำคัญของตัวแปรคุณลักษณะแต่ะตัว (Feature Importance)

เรียงจากมากไปน้อยใน Random Forest

Features	Feature Importance
ST slope	0.193879
exercise angina	0.144869
chest pain type	0.120106
oldpeak	0.107030
max heart rate	0.102770
age	0.101393
cholesterol	0.082878
resting bp s	0.074385
sex	0.035551
resting ecg	0.029120
fasting blood sugar	0.008020
Total	**1.000000**

4.6.2 ชุดของตัวแปรและพารามิเตอร์ที่ดีที่สุด

4.6.2.1 ชุดของตัวแปร

ชุดของตัวแปรที่ใช้ทั้ง 10 ชุด โดยจะคัดตัวแปรที่มีน้ำหนักความสำคัญของตัวแปร (Feature Importance) น้อยที่สุดออก ได้แก่

X_1 = ['age', 'sex', 'chest pain type', 'resting bp s', 'cholesterol', 'fasting blood sugar', 'resting ecg', 'max heart rate', 'exercise angina', 'oldpeak', 'ST slope']

X_2 = ['age', 'sex', 'chest pain type', 'resting bp s', 'cholesterol', 'resting ecg', 'max heart rate', 'exercise angina', 'oldpeak', 'ST slope']

X_3 = ['age', 'chest pain type', 'resting bp s', 'cholesterol', 'max heart rate', 'exercise angina', 'oldpeak', 'ST slope']

X_4 = ['age', 'sex', 'chest pain type', 'resting bp s', 'cholesterol', 'max heart rate', 'exercise angina', 'oldpeak', 'ST slope']

X_5 = ['age', 'chest pain type', 'cholesterol', 'max heart rate', 'exercise angina', 'oldpeak', 'ST slope']

X_6 = ['age', 'chest pain type', 'max heart rate', 'exercise angina', 'oldpeak', 'ST slope']

X_7 = ['chest pain type','max heart rate', 'exercise angina', 'oldpeak', 'ST slope']

X_8 = ['chest pain type', 'exercise angina', 'oldpeak', 'ST slope']

X_9 = ['chest pain type', 'exercise angina', 'ST slope']

X_10 = ['exercise angina', 'ST slope']

4.6.2.2 แบบจำลองที่ดีที่สุดจากแต่ละชุดของตัวแปร

ตารางที่ 4.18 แบบจำลองที่ดีที่สุดในแต่ละชุดของตัวแปรใน Random Forest

ชุดของตัวแปร	Model (random_state = 1000)	Accuracy Score
X_1	criterion= 'gini', max_depth = 12, n_estimators = 17	0.908497
X_2	criterion= 'gini', max_depth = 9, n_estimators = 7	0.924837
X_3	criterion='entropy', max_depth = 11, n_estimators = 17	0.918301
X_4	criterion='gini', max_depth = 11, n_estimators = 17	0.905229
X_5	criterion='entropy', max_depth = 11, n_estimators = 19	0.911765
X_6	criterion='entropy', max_depth = 11, n_estimators = 13	0.898693
X_7	criterion='gini', max_depth = 12, n_estimators = 17	0.885621

ชุดของตัวแปร	Model (random_state = 1000)	Accuracy Score
X_8	criterion='entropy', max_depth = 5, n_estimators = 5	0.843137
X_9	criterion='entropy', max_depth = 6, n_estimators = 5	0.823529
X_10	criterion='entropy', max_depth = 3, n_estimators = 7	0.784314

จากตารางที่ 4.18 แสดงให้เห็นว่า ชุดของตัวแปรที่ดีที่สุดในแบบจำลองที่ดีที่สุดคือชุดข้อมูลที่ X_2 ซึ่งประกอบด้วยตัวแปร age, sex, chest pain type, resting bp s, cholesterol, resting ecg, max heart rate, exercise angina, oldpeak และ ST slope โดยมี Accuracy Score เท่ากับ 0.924837 ซึ่งเป็น RandomForestClassifier ที่มีพารามิเตอร์

1. criterion = ‘gini’
2. max_depth = 9
3. n_estimators = 7
4. random_state = 1000

4.6.3 ประสิทธิภาพของแบบจำลอง

ตารางที่ 4.19 ประสิทธิภาพของแบบจำลอง ทดสอบด้วย Training Data

Train	Precision	Recall	F1-score	Support
Accuracy		0.9803		711
0	0.9850	0.9801	0.9825	402
1	0.9743	0.9806	0.9774	309
Macro avg	0.9796	0.9803	0.9800	711
Weighted avg	0.9803	0.9803	0.9803	711

Accuracy = 0.9803 หมายความว่า จำนวนการคาดการณ์ที่ถูกต้องทุกประเภท คิดเป็นร้อยละ 98.03 ของจำนวนการคาดการณ์ทั้งหมด

Precision

[Target = 0] = 0.9850 หมายความว่า ในจำนวนคนที่แบบจำลองคาดการณ์ว่าไม่เป็นโรคหัวใจทั้งหมด มีคนที่ไม่ได้เป็นโรคหัวใจจริงร้อยละ 98.50

[Target = 1] = 0.9743 หมายความว่า ในจำนวนคนที่แบบจำลองคาดการณ์ว่าเป็นโรคหัวใจทั้งหมด มีคนที่เป็นโรคหัวใจจริงร้อยละ 97.43

Recall

[Target = 0] = 0.9801 หมายความ ในจำนวนคนที่ไม่เป็นโรคหัวใจจริงทั้งหมด แบบจำลองคาดการณ์ว่าไม่เป็นโรคหัวใจร้อยละ 98.01

[Target = 1] = 0.9806 หมายความว่า ในจำนวนคนที่เป็นโรคหัวใจจริงทั้งหมด แบบจำลองคาดการณ์ว่าเป็นโรคหัวใจร้อยละ 98.06

F1-score

[Target = 0] = 0.9825 หมายความว่า เมื่อพิจารณาคนที่ไม่เป็นโรคหัวใจจริง ค่าเฉลี่ยฮาร์โมนิคของ Precision และ Recall คือร้อยละ 98.25

[Target = 1] = 0.9774 หมายความว่า เมื่อพิจารณาคนที่เป็นโรคหัวใจจริง ค่าเฉลี่ยฮาร์โมนิคของ Precision และ Recall คือร้อยละ 97.74

ตารางที่ 4.20 ประสิทธิภาพของแบบจำลอง ทดสอบด้วย Testing Data

Train	Precision	Recall	F1-score	Support
Accuracy		0.9248		306
0	0.9028	0.9353	0.9187	139
1	0.9444	0.9162	0.9301	167
Macro avg	0.9236	0.9257	0.9244	306
Weighted avg	0.9255	0.9248	0.9249	306

Accuracy = 0.9248 หมายความว่า จำนวนการคาดการณ์ที่ถูกต้องทุกประเภท คิดเป็นร้อยละ 92.48 ของจำนวนการคาดการณ์ทั้งหมด

Precision

[Target = 0] = 0.9028 หมายความว่า ในจำนวนคนที่แบบจำลองคาดการณ์ว่าไม่เป็นโรคหัวใจทั้งหมด มีคนที่ไม่ได้เป็นโรคหัวใจจริงร้อยละ 90.28

[Target = 1] = 0.9444 หมายความว่า ในจำนวนคนที่แบบจำลองคาดการณ์ว่าเป็นโรคหัวใจทั้งหมด มีคนที่เป็นโรคหัวใจจริงร้อยละ 94.44

Recall

[Target = 0] = 0.9353 หมายความว่า ในจำนวนคนที่ไม่เป็นโรคหัวใจจริงทั้งหมด แบบจำลองคาดการณ์ว่าไม่เป็นโรคหัวใจร้อยละ 93.53

[Target = 1] = 0.9162 หมายความว่า ในจำนวนคนที่เป็นโรคหัวใจจริงทั้งหมด แบบจำลองคาดการณ์ว่าเป็นโรคหัวใจร้อยละ 91.62

F1-score

[Target = 0] = 0.9187 หมายความว่า เมื่อพิจารณาคนที่ไม่เป็นโรคหัวใจจริง ค่าเฉลี่ยฮาร์โมนิคของ Precision และ Recall คือร้อยละ 91.87

[Target = 1] = 0.9301 หมายความว่า เมื่อพิจารณาคนที่เป็นโรคหัวใจจริง ค่าเฉลี่ยฮาร์โมนิคของ Precision และ Recall คือร้อยละ 93.01

4.7 สรุป และอภิปรายผล

จากแบบจำลองทั้ง 2 แบบ ได้ทำการคัดเลือกเฉพาะแบบจำลองที่ดีที่สุดแล้ว โดยนำประสิทธิภาพของทั้ง 2 แบบจำลองมาเปรียบเทียบกัน ด้วยการใช้เกณฑ์ Precision, Recall, F1-score และ Accuracy เพื่อหาแบบจำลองที่มีประสิทธิภาพดีที่สุด ซึ่งคือแบบจำลองป่าไม้สุ่ม (Random Forest Model) ที่มีพารามิเตอร์คือ criterion= 'gini', max_depth = 9 และ n_estimators = 7

บทที่ 5

สรุปผลการวิจัย และข้อเสนอแนะ

5.1 สรุปผลการวิจัย

จากการวิเคราะห์ข้อมูลเชิงคุณภาพของกลุ่มตัวอย่างทั้งหมด กลุ่มตัวอย่างส่วนใหญ่ไม่เป็นโรคหัวใจ และส่วนใหญ่จะเป็นเพศชาย ซึ่งมีจำนวน 747 คน รวมถึงไม่มีอาการเจ็บปวดขณะออกกำลัง ซึ่งมีจำนวน 640 คน โดยเป็นคนที่ไม่มีอาการเจ็บหน้าอกเลย 499 คน ขณะที่ตัวอย่างที่มีครบ 3 ลักษณะของอาการเจ็บหน้าอกแบบ Typical angina มีจำนวน 61 คน ในด้านของค่าน้ำตาลในเลือด ส่วนใหญ่มีค่าน้ำตาลกลูโคสต่ำกว่าหรือเท่ากับ 120 มิลลิกรัมต่อเดซิลิตร ซึ่งมีจำนวน 853 คน นอกจากนี้กลุ่มตัวอย่างส่วนใหญ่เป็นคนที่มีคลื่นไฟฟ้าหัวใจขณะพักปกติ ซึ่งมีจำนวน 576 คน และมีค่า ST segment มีความชันเป็นบวก ซึ่งมีจำนวน 480 คน

จากผลการวิเคราะห์ข้อมูลทั้งหมด จะเห็นได้ว่าค่าเฉลี่ยของตัวแปรเชิงปริมาณต่าง ๆ ของกลุ่มที่ไม่เป็นโรคหัวใจ (target = 0) จะมีอายุเฉลี่ยอยู่ที่ 50.87 ปี ความดันโลหิตขณะพัก 129.77 มิลลิเมตรปรอท ระดับคอเลสเตอรอลในเลือด 240.22 มิลลิกรัมต่อเดซิลิตร อัตราการเต้นของหัวใจสูงสุด 151.65 ครั้งต่อนาที และมีระดับการยกตัวของ ST segment เทียบกับ Isoelectric Line 0.46 มิลลิเมตร ในขณะที่กลุ่มที่เป็นโรคหัวใจ (target = 1) เป็นดังนี้ อายุเฉลี่ยอยู่ที่ 56.02 ปี ความดันโลหิตขณะพัก 135.72 มิลลิเมตรปรอท ระดับคอเลสเตอรอลในเลือด 252.47 มิลลิกรัมต่อเดซิลิตร อัตราการเต้นของหัวใจสูงสุด 132.62 ครั้งต่อนาที และมีระดับการยกตัวของ ST segment เทียบกับ Isoelectric Line 1.49 มิลลิเมตร โดยจะสังเกตได้ว่าทุกตัวแปรเชิงปริมาณ กลุ่มที่เป็นโรคหัวใจจะมีค่ามากกว่ากลุ่มที่ไม่เป็นโรคหัวใจ ยกเว้นอัตราการเต้นของหัวใจสูงสุด

ในแบบจำลองต้นไม้ตัดสินใจ (Decision Tree Model) ค่า Accuracy สูงสุดที่ได้จากข้อมูลชุด Testing Data คือ 0.9052 โดยใช้ Test size = 0.3 Random state = 1000 โดยมีพารามิเตอร์ที่ดีที่สุดคือ criterion= 'gini' และ max_depth = 12 โดยมี feature ทั้งหมด 10 ตัว ได้แก่ อายุ (age) เพศ (sex) ลักษณะอาการเจ็บหน้าอก (chest pain type) ความดันโลหิตขณะพัก (resting bp s) ระดับของคอเลสเตอรอลในเลือด (cholesterol) ระดับน้ำตาลในเลือดในช่วง 2-3 วันที่ผ่านมา โดยงดอาหาร 8 ชั่วโมงก่อนตรวจ (fasting blood sugar) อัตราการเต้นของหัวใจสูงสุด (max heart rate) อาการเจ็บหน้าอกขณะออกกำลังกาย (exercise angina) ระดับการยกตัวของ ST segment เทียบกับ Isoelectric Line (oldpeak) และลักษณะความชันของ ST segment (ST slope) โดยที่ค่า Precision score สำหรับแต่ละกลุ่มคือ {[target = 0] : 0.8929, [target = 1] = 0.9157} ค่า Recall สำหรับแต่ละกลุ่มคือ {[target = 0] : 0.8993, [target = 1] = 0.9192} ค่า F1-score สำหรับแต่ละกลุ่มคือ {[target = 0] : 0.8961, [target = 1] = 0.9129}

ในแบบจำลองป่าไม้สุ่ม (Random Forest Model) ค่า Accuracy สูงสุดที่ได้จากข้อมูลชุด Testing Data คือ 0.9248 โดยใช้ Test size = 0.3 Random state = 1000 โดยมีพารามิเตอร์ที่ดีที่สุดคือ criterion= 'gini', max_depth = 9 และ n_estimators = 7 โดยมี feature ทั้งหมด 10 ตัว ได้แก่ อายุ (age) เพศ (sex) ลักษณะอาการเจ็บหน้าอก (chest pain type) ความดันโลหิตขณะพัก (resting bp s) ระดับของคอเลสเตอรอลในเลือด (cholesterol) คลื่นไฟฟ้าหัวใจขณะพัก (resting ecg) อัตราการเต้นของหัวใจสูงสุด (max heart rate) อาการเจ็บหน้าอกขณะออกกำลังกาย (exercise angina) ระดับการยกตัวของ ST segment เทียบกับ Isoelectric Line (oldpeak) และลักษณะความชันของ ST segment (ST slope) โดยที่ค่า Precision score สำหรับแต่ละกลุ่มคือ {[target = 0] : 0.9028, [target = 1] = 0.9444} ค่า Recall สำหรับแต่ละกลุ่มคือ {[target = 0] : 0.9353, [target = 1] = 0.9162} ค่า F1-score สำหรับแต่ละกลุ่มคือ {[target = 0] : 0.9187, [target = 1] = 0.9301}

เมื่อนำแบบจำลองทั้งสองประเภทมาเปรียบเทียบประสิทธิภาพด้วยตารางด้านล่าง จะได้ผลลัพธ์ดังนี้

ตารางที่ 5.1 การเปรียบเทียบประสิทธิภาพของแบบจำลองป่าไม้สุ่ม และแบบจำลองต้นไม้ตัดสินใจ

ค่าประสิทธิภาพจากชุดข้อมูล		**แบบจำลองต้นไม้ตัดสินใจ (Decision Tree Model)**	**แบบจำลองป่าไม้สุ่ม (Random Forest Model)**
Precision	**0**	**0.9052**	**0.9028**
	1	**0.9157**	**0.9444**
Recall	**0**	**0.8993**	**0.9353**
	1	**0.9192**	**0.9162**
F1-score	**0**	**0.8961**	**0.9187**
	1	**0.9129**	**0.9301**
Accuracy		**0.9052**	**0.9284**

จากตารางจะเห็นได้ว่าค่าประสิทธิภาพของแบบจำลอง Precision, Recall, F1-score และ Accuracy ในกลุ่ม 0 และ 1 แบบจำลองป่าไม้สุ่ม (Random Forest Model) ส่วนใหญ่มีค่าประสิทธิภาพสูงกว่า จึงสามารถสรุปได้ว่าแบบจำลองป่าไม้สุ่ม (Random Forest Model) เป็นแบบจำลองที่ดีที่สุดโดยที่ใช้ Test size = 0.3 และมีตัวแปรอิสระ 10 ตัว ได้แก่ age, sex, chest pain type, resting bp s, cholesterol, resting ecg, max heart rate, exercise angina, oldpeak และ ST slope ตัวแปรตามคือ Target ว่าเป็นโรคหัวใจหรือไม่ โดยใช้พารามิเตอร์ดังนี้ criterion= 'gini', max_depth = 9 และ n_estimators = 7

5.2 ข้อเสนอแนะ

จะเห็นได้ว่าการสร้างแบบจำลองป่าไม้สุ่ม (Random Forest) และแบบจำลองต้นไม้ตัดสินใจ (Decision Tree) มีประสิทธิภาพในเกณฑ์ Accuracy Score ในการจำแนกผู้ที่ไม่เป็นโรคหัวใจ (target = 0) และผู้ที่เป็นโรคหัวใจ (target = 1) ดีมาก แต่กลุ่มตัวอย่างที่นำมาใช้ศึกษานั้นยังถือว่าไม่มีประสิทธิภาพพอ เนื่องจากกลุ่มตัวอย่างส่วนใหญ่เป็นเพศชาย และมีค่าคอเลสเตอรอลสูง อีกทั้งยังมีอัตราการเต้นของหัวใจสูง จึงทำให้ผลลัพธ์ที่ได้มีค่าความแม่นยำมาก หากใช้กับกลุ่มคนที่เป็นเพศหญิง มีค่าคอเลสเตอรอลปกติ หรือมีอัตราการเต้นของหัวใจปกติ ก็อาจทำให้ค่า Accuracy Score ลดลงได้ แต่เนื่องจากผู้ศึกษามีระยะเวลาในการหาข้อมูลที่จำกัด ดังนั้นหากจะให้โมเดลนี้ดีขึ้น ควรจะหากลุ่มตัวอย่างเพิ่ม และคละตัวอย่างให้มีความหลากหลายมากขึ้น

นอกจากนี้ หากมีผู้ที่ต้องการใช้แบบจำลองเพื่อใช้ในการจำแนกประเภทเฉพาะทาง สามารถเลือกปรับเกณฑ์ที่เหมาะสมกับแบบจำลองนั้น ๆ ได้ เช่น หากโรงพยาบาลต้องการคัดเลือกผู้ที่มีอาการโรคหัวใจ อาจปรับเกณฑ์การคัดเลือกเป็น Recall Score เพื่อที่จะให้แบบจำลองคัดกรองผู้ที่มีอาการโรคหัวใจออกมาได้มากที่สุดจากกลุ่มตัวอย่างทั้งหมด

บรรณานุกรม

Geeksforgeeks. (2567). **Confusion Matrix in Machine Learning**. สืบค้น 28 พฤษภาคม 2567, จาก https://www.geeksforgeeks.org/confusion-matrix-machine-learning/

IBM. (ม.ป.ป.). **What is ML?.** สืบค้น 26 พฤษภาคม 2567, จาก https://www.ibm.com/topics/machine-learning

IBM. (ม.ป.ป.). **What is random forest?** สืบค้น 28 พฤษภาคม 2567, จาก https://www.ibm.com/topics/random-forest

Mexwell. (2567). **Heart Disease Dataset**. สืบค้น 25 พฤษภาคม 2567, จาก https://www.kaggle.com/datasets/mexwell/heart-disease-dataset/discussion

Phanpaporn. (ม.ป.ป.). **ทำความรู้จัก "Linear Regression" Algorithm ที่คนทำ Machine Learning ยังไงก็ต้องได้ใช้!** สืบค้น 27 พฤษภาคม 2567, จาก https://www.borntodev.com/2021/08/26/

UC Berkeley. (2565). **What Is Machine Learning (ML)?** สืบค้น 26 พฤษภาคม 2567, จาก https://ischoolonline.berkeley.edu/blog/what-is-machine-learning/

Witchapong Daroontham. (2561). **รู้จัก Decision Tree, Random Forest, และ XGBoost!!! — PART 1.** สืบค้น 28 พฤษภาคม 2567, จาก https://medium.com/@witchapongdaroontham

Zoumana Kelta. (2565). Classification in Machine Learning: An Introduction. สืบค้น 27 พฤษภาคม 2567, จาก https://www.datacamp.com/blog/classification-machine-learning

กิตติมศักดิ์ ในจิต. (ม.ป.ป.). **การ Train และ Test ใน Machine Learning ด้วย Python**. สืบค้น 26 พฤษภาคม 2567, จาก https://kittimasak.com/train-test-machine-learning-python/

กองโรคไม่ติดต่อ. (2567). **รายงานประจำปี 2566 กองโรคไม่ติดต่อ**. สืบค้น 25 พฤษภาคม 2567, จาก https://ddc.moph.go.th/dncd/news.php?news=41556&deptcode=

ชลธร วงศ์รัศมี. (2561). ลดคนล้นโรงพยาบาล ด้วยความคิดเชิงระบบ กับ บวรศม ลีระพันธ์. สืบค้น 25 พฤษภาคม 2567, จาก https://www.the101.world/system_thinking_health/

ณัฐรฐนนท์ กานต์รวีกุลธนา. (2019). **Overfitting กับการแก้ปัญหา**. สืบค้น 28 พฤษภาคม 2567, จาก https://medium.com/@natratanonkanraweekultana/overfitting-

ไพศาล บุญศิริคำชัย. (2564). **โรคหัวใจ (Heart Disease)**. สืบค้น 25 พฤษภาคม 2567, จาก https://www.medparkhospital.com/disease-and-treatment/heart-disease

ไพสิฐ วิสัตยรักษ์. (2023). **Regression คืออะไร มาเรียนรู้แบบง่าย ๆ สไตล์เด็กวิทย์คอมกัน (ตอนที่ 1)**. สืบค้น 27 พฤษภาคม 2567, จาก https://medium.com/@2pm.tayyoshi/regression

โรงพยาบาลรามคำแหง. (2567). ความดันโลหิตสูงที่ทำให้เป็นโรคหัวใจ. สืบค้น 25 พฤษภาคม 2567, จาก https://www.ram-hosp.co.th/news_detail/240

โรงพยาบาลศิครินทร์ กรุงเทพ. (ม.ป.ป). **โรคความดันโลหิตสูง ภัยเงียบต่อหลอดเลือดและหัวใจ**. สืบค้น 25 พฤษภาคม 2567, จาก https://www.sikarin.com/health

โรงพยาบาลสินแพทย์ รามอินทรา. (2564). ทำไม...? ต้องวัดความดันโลหิตที่บ้าน. สืบค้น 25 พฤษภาคม 2567, จากhttps://www.synphaet.co.th

สมศักดิ์ ชุณหรัศมิ์. (2566). **ความจริงนโยบายสาธารณสุข โจทย์ท้าทายรัฐบาลหลังเลือกตั้ง: ตอนที่ 1 ความแออัดที่โรงพยาบาล.** สืบค้น 25 พฤษภาคม 2567, จาก https://www.the101.world/healthcare-policy-1/