# ประสิทธิภาพการทำงานของโมเดลต้นไม้ตัดสินใจสำหรับการพยากรณ์ข้อมูลด้วยชุดข้อมูลการพยากรณ์โรคหัวใจ

# Performance of the Decision Tree Model for Data Prediction Using a Heart Disease Prediction Dataset

**สันทัด พรหมแท่น[1], ณธัชสร จุติสงขลา[2], ธัชตะวัน ชนะกูล[3]**

*[1]สาขาวิชาเทคโนโลยีสารสนเทศ, คณะวิทยาศาสตร์และเทคโนโลยี, มหาวิทยาลัยหาดใหญ่, santad@hu.ac.th*

*[2]สาขาวิชาเทคโนโลยีสารสนเทศ, คณะวิทยาศาสตร์และเทคโนโลยี, มหาวิทยาลัยหาดใหญ่, nathatsorn@hu.ac.th*

*[3]สาขาวิชาระบบสารสนเทศทางธุรกิจ, คณะบริหารธุรกิจ, มหาวิทยาลัยหาดใหญ่, tawan_chana@hu.ac.th*

**บทคัดย่อ**

โรคหัวใจเป็นหนึ่งในสาเหตุหลักที่ทำให้เกิดการเสียชีวิตทั่วโลกและมีแนวโน้มเพิ่มสูงขึ้น ส่งผลกระทบต่อคุณภาพชีวิตของผู้ป่วยและระบบสาธารณสุข ดังนั้น การพัฒนาโมเดลพยากรณ์โรคหัวใจด้วยเทคนิคการเรียนรู้ของเครื่อง เช่น ต้นไม้ตัดสินใจ (Decision Tree) จึงเป็นแนวทางสำคัญในการตรวจวินิจฉัยอย่างรวดเร็วและแม่นยำ การวิจัยครั้งนี้มีวัตถุประสงค์ 1) เพื่อพัฒนาโมเดลต้นไม้ตัดสินใจสำหรับพยากรณ์ผู้ป่วยโรคหัวใจ 2) เพื่อประเมินประสิทธิภาพการทำงานของโมเดลต้นไม้ตัดสินใจ โดยใช้วิธีดำเนินการวิจัยตามกระบวนการพื้นฐานของวิทยาศาสตร์ข้อมูลทั้งหมด 5 ขั้นตอน ประกอบด้วย การเก็บรวบรวมข้อมูล การสำรวจข้อมูล การเตรียมข้อมูล การเลือกโมเดล และการประเมินประสิทธิภาพการทำงานของโมเดล

ผลการวิจัยพบว่า ประสิทธิภาพการทำงานของโมเดลต้นไม้ตัดสินใจสำหรับการพยากรณ์ข้อมูลด้วยชุดข้อมูลการพยากรณ์โรคหัวใจ พบว่าโมเดลต้นไม้ตัดสินใจมีค่าความถูกต้อง (Accuracy) เท่ากับ 98.54%, ค่าความแม่นยำ (Precision) เท่ากับ 100%, ค่าความระลึก (Recall) เท่ากับ 97.09% และค่า F1 (F1-score) เท่ากับ 98.52%

***คำหลัก**:* ต้นไม้ตัดสินใจ โมเดลพยากรณ์ การประเมินประสิทธิภาพ โรคหัวใจ

**Abstract**

Heart disease is one of the leading causes of death worldwide, and its incidence continues to rise. This condition significantly impacts patients' quality of life as well as the public healthcare system. Therefore, developing predictive models for heart disease using machine learning techniques, such as Decision Trees, has become an essential approach for rapid and accurate diagnosis.

The objectives of this research are: (1) to develop a Decision Tree model for predicting heart disease, and (2) to evaluate the performance of the Decision Tree model. The research was conducted following the fundamental five-step data science process, which includes data collection, data exploration, data preparation, model selection, and model performance evaluation.

The research findings revealed that the performance of the Decision Tree model for heart disease prediction demonstrated high effectiveness. The model achieved an accuracy of 98.54%, a precision of 100%, a recall of 97.09%, and an F1-score of 98.52%.

***Keywords:*** Decision Tree, Prediction Model, Performance Evaluation, Heart Disease

**ความเป็นมาและความสำคัญของปัญหา**

โรคหัวใจเป็นโรคที่เกิดจากความผิดปกติของหัวใจและระบบไหลเวียนโลหิต หัวใจเป็นอวัยวะสำคัญที่ทำหน้าที่สูบฉีดเลือดไปเลี้ยงทั่วร่างกาย หากหัวใจทำงานผิดปกติจะส่งผลกระทบต่อสุขภาพโดยรวม โรคหัวใจสามารถเกิดได้จากหลายปัจจัย เช่น พันธุกรรม การใช้ชีวิตที่ไม่เหมาะสม และโรคเรื้อรัง หนึ่งในโรคหัวใจที่พบบ่อยคือโรคหลอดเลือดหัวใจตีบ ซึ่งเกิดจากไขมันสะสมในหลอดเลือด ทำให้เลือดไปเลี้ยงหัวใจไม่เพียงพอ หากรุนแรงอาจทำให้เกิดหัวใจวายได้ อาการของโรคหัวใจอาจเริ่มจากการเจ็บหน้าอก เหนื่อยง่าย ใจสั่น หรือหายใจลำบาก หากปล่อยทิ้งไว้อาจนำไปสู่ภาวะแทรกซ้อนที่อันตราย ปัจจัยเสี่ยงที่ทำให้เกิดโรคหัวใจ ได้แก่ การรับประทานอาหารที่มีไขมันสูง ขาดการออกกำลังกาย สูบบุหรี่ ดื่มแอลกอฮอล์ ความเครียด เป็นต้น การควบคุมปัจจัยเสี่ยงเหล่านี้สามารถช่วยลดโอกาสการเกิดโรคหัวใจได้ วิธีป้องกันที่สำคัญคือการเลือกรับประทานอาหารที่มีประโยชน์ ออกกำลังกายเป็นประจำ และดูแลสุขภาพจิต การตรวจโรคหัวใจเป็นประจำช่วยให้พบความผิดปกติได้เร็วขึ้น หากตรวจพบโรคหัวใจในระยะเริ่มต้น แพทย์สามารถให้คำแนะนำและรักษาได้อย่างเหมาะสม การรักษาโรคหัวใจมีหลายวิธี ตั้งแต่การปรับเปลี่ยนพฤติกรรม การใช้ยา ไปจนถึงการผ่าตัดขึ้นอยู่กับความรุนแรงของโรค แม้ว่าโรคหัวใจจะเป็นโรคร้ายแรง แต่สามารถป้องกันได้หากมีการดูแลสุขภาพอย่างถูกต้อง การมีสุขภาพหัวใจที่ดีช่วยให้มนุษย์มีชีวิตที่ยืนยาวและมีคุณภาพชีวิตที่ดีขึ้น

สมาพันธ์หัวใจโลก (World Heart Federation, WHF) ได้เปิดเผยข้อมูลสถิติปี 2567 เกี่ยวกับโรคหัวใจและหลอดเลือด พบว่า มีผู้เสียชีวิตจากโรคหัวใจและหลอดเลือดมากที่สุดเป็นอันดับ 1 ของโลก โดยมีจำนวนผู้เสียชีวิตมากกว่า 20.5 ล้านคนต่อปี หรือเฉลี่ยมากกว่า 39 คนต่อนาที ที่ต้องเสียชีวิตลงจากโรคนี้อย่างเฉียบพลันและหลีกเลี่ยงไม่ได้ ร้อยละ 85 ของการเสียชีวิตเกิดจากอาการหัวใจวายเฉียบพลันและโรคหลอดเลือดสมอง ซึ่งอาจเกิดขึ้นได้ทุกเมื่อโดยไม่มีสัญญาณเตือนล่วงหน้า สถานการณ์นี้ไม่เพียงส่งผลกระทบในระดับโลก แต่ยังเป็นภัยคุกคามร้ายแรงต่อสุขภาพของคนไทย จากข้อมูลของระบบคลังข้อมูลด้านการแพทย์และสุขภาพ กระทรวงสาธารณสุข (Health Data Center) ปี 2566 พบว่าประเทศไทยมีผู้ป่วยสะสมด้วยโรคหัวใจและหลอดเลือด

มากกว่า 250,000 ราย และมีจำนวนผู้เสียชีวิตมากถึง 40,000 รายต่อปี หรือเฉลี่ยชั่วโมงละ 5 คน ซึ่งหมายความว่า ทุก 12 นาที จะมีคนไทย 1 คนต้องเสียชีวิตลงจากโรคหัวใจ โรคหัวใจและหลอดเลือดไม่ได้เป็นเพียงโรคของผู้สูงอายุอีกต่อไป แต่กำลังคุกคามคนทุกช่วงอายุและวัยทำงานที่มีพฤติกรรมเสี่ยงโดยไม่รู้ตัว หากไม่มีมาตรการป้องกันที่เข้มงวด จำนวนผู้เสียชีวิตจะเพิ่มขึ้นอย่างน่าตกใจ และอาจทำให้ระบบสาธารณสุขต้องเผชิญกับวิกฤตครั้งใหญ่ในอนาคต

เนื่องจากโรคหัวใจเป็นสาเหตุการเสียชีวิตอันดับหนึ่งของโลกและมีแนวโน้มเพิ่มขึ้นอย่างต่อเนื่อง ส่งผลกระทบอย่างรุนแรงต่อคุณภาพชีวิตของผู้ป่วย ครอบครัว และระบบสาธารณสุข ดังนั้น การพัฒนางานวิจัยเพื่อพยากรณ์โรคหัวใจด้วยโมเดลต้นไม้ตัดสินใจ ได้รับความสนใจเนื่องจากความสามารถในการตีความผลลัพธ์ได้ง่าย, การจัดการกับข้อมูลที่ไม่สมบูรณ์ได้ดี, และไม่ต้องการการแปลงข้อมูลล่วงหน้ามากนัก โดยงานวิจัยก่อนหน้านี้แสดงให้เห็นว่า Decision Tree สามารถประเมินความเสี่ยงจากโรคหัวใจได้อย่างมีประสิทธิภาพและช่วยในการตัดสินใจทางการแพทย์ได้อย่างแม่นยำ, ทำให้เป็นเครื่องมือที่มีศักยภาพในการช่วยในการพยากรณ์โรคหัวใจในผู้ป่วยได้อย่างเหมาะสมและมีประสิทธิผล จึงทำให้เป็นแนวทางสำคัญในการช่วยให้การตรวจวินิจฉัยเป็นไปอย่างรวดเร็วและแม่นยำ ด้วยการนำข้อมูลการพยากรณ์โรคหัวใจจากชุดข้อมูลเปิดใน Kaggle มาพัฒนาโมเดลต้นไม้ตัดสินใจเพื่อพยากรณ์การเกิดโรคหัวใจและทำการประเมินประสิทธิภาพการทำงานของโมเดล ช่วยให้แพทย์สามารถตัดสินใจรักษาได้อย่างมีประสิทธิภาพ นอกจากนี้ การพยากรณ์โรคหัวใจยังช่วยลดภาระของบุคลากรทางการแพทย์ ลดค่าใช้จ่ายในการรักษา และเพิ่มโอกาสรอดชีวิตให้กับผู้ป่วย การพัฒนาเทคนิคดังกล่าวไม่เพียงช่วยให้เกิดการดูแลสุขภาพแบบเชิงรุก แต่ยังสามารถนำไปใช้ในการวางแผนนโยบายสาธารณสุขเพื่อลดอัตราการเสียชีวิตในระยะยาว ดังนั้น งานวิจัยชิ้นนี้จึงมีความสำคัญอย่างยิ่งในการสร้างระบบสาธารณสุขที่มีประสิทธิภาพและยั่งยืนในอนาคต

## วัตถุประสงค์

1. เพื่อพัฒนาโมเดลต้นไม้ตัดสินใจสำหรับพยากรณ์ผู้ป่วยโรคหัวใจ
2. เพื่อประเมินประสิทธิภาพการทำงานของโมเดลต้นไม้ตัดสินใจ

## ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

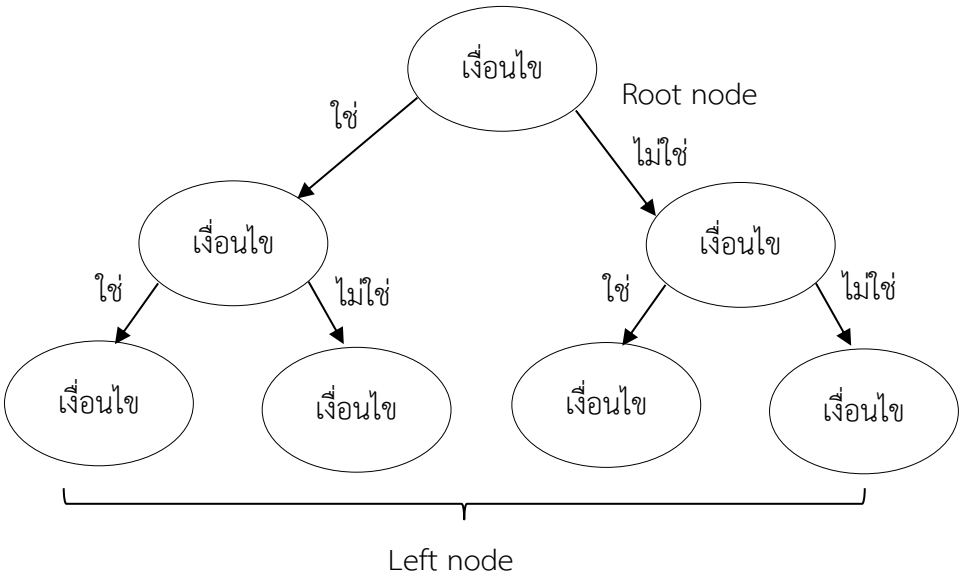
### 1. ทฤษฎีที่เกี่ยวข้อง

### ต้นไม้ตัดสินใจ (Decision Tree)

เป็นหนึ่งในอัลกอริทึมของการเรียนรู้ของเครื่อง (Machine Learning) ที่ใช้สำหรับการตัดสินใจหรือการจำแนกประเภทข้อมูลโดยวิธีการสร้างโครงสร้างที่คล้ายกับต้นไม้ (Tree structure) ซึ่งประกอบด้วยโหนด (Nodes) และ กิ่ง (Branches) ที่แสดงถึงการตัดสินใจหรือการเลือกเส้นทางตามข้อมูลที่ได้รับมา (ศุภจิตรี บุญอำนวย, 2562) การทำงานของต้นไม้ตัดสินใจมีขั้นตอน ดังนี้

1. การเลือกคุณลักษณะ (Features) เป็นขั้นตอนที่ต้นไม้ตัดสินใจจะเลือกคุณลักษณะ (feature) ที่สำคัญที่สุดจากข้อมูลที่มี เพื่อใช้สำหรับการแบ่งข้อมูลออกเป็นกลุ่มย่อย หรือเรียกว่าการแบ่งโหนด (Node splitting) ซึ่งการเลือกคุณลักษณะที่เหมาะสมจะช่วยให้ต้นไม้ตัดสินใจมีความแม่นยำมากขึ้น
2. การแบ่งกลุ่มข้อมูล (Data Splitting) เป็นขั้นตอนของการแบ่งข้อมูลออกเป็นกลุ่มตามเงื่อนไขที่กำหนดไว้ในแต่ละโหนด เช่น หากเป็นข้อมูลที่มีคุณลักษณะของตัวแปรที่เป็นตัวเลข อาจจะแบ่งข้อมูลตามช่วงของตัวเลข หรือถ้าเป็นข้อมูลประเภทหมวดหมู่ ก็จะแบ่งออกตามค่าของหมวดหมู่
3. การทำซ้ำ (Recursion) เป็นขั้นตอนที่ระบบจะทำการทดสอบเลือกคุณลักษณะอื่น ๆ เพื่อแบ่งข้อมูลในกลุ่มย่อยนั้นต่อไป จนกว่าข้อมูลจะถูกแบ่งออกจนหมด หรือจนกว่าจะถึงเงื่อนไขที่กำหนด
4. การตัดสินใจสุดท้าย (Leaf Nodes) เป็นขั้นตอนที่โหนดสุดท้ายที่ไม่สามารถแบ่งข้อมูลเพิ่มเติมได้อีก จะกลายเป็นโหนดใบไม้ (Leaf node) ซึ่งจะเป็นคำตอบหรือผลลัพธ์จากการทำนาย เช่น ในกรณีของการจำแนกประเภท (Classification) โหนดใบไม้จะบอกว่าข้อมูลนั้นอยู่ในกลุ่มใด หรือในกรณีของการพยากรณ์จะเป็นค่าที่ถูกคาดการณ์ไว้ล่วงหน้า

   แสดงโครงสร้างของต้นไม้ตัดสินใจดังภาพที่ 1




ภาพ 1 แสดงโครงสร้างของต้นไม้ตัดสินใจ

## 2. งานวิจัยที่เกี่ยวข้อง

สุรวัชร ศรีเปารยะ และและสายชล สินสมบูรณ์ทอง (2560) ได้นำข้อมูลโรคไตเรื้อรังมาวิเคราะห์ผ่านอัลกอรึทึมด้วยวิธีความใกล้เคียงกันมากที่สุด (K-nearest neighbor) วิธีต้นไม้ตัดสินใจ (Decision tree) วิธีโครงข่ายประสาทเทียม (Artificial neural network) วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support vector machine), วิธีฐานกฎ (Rule-based), วิธีการถดถอยลอจิสติก (Logistic regression), และวิธีนาอีฟเบย (Naïve Bayes) โดย

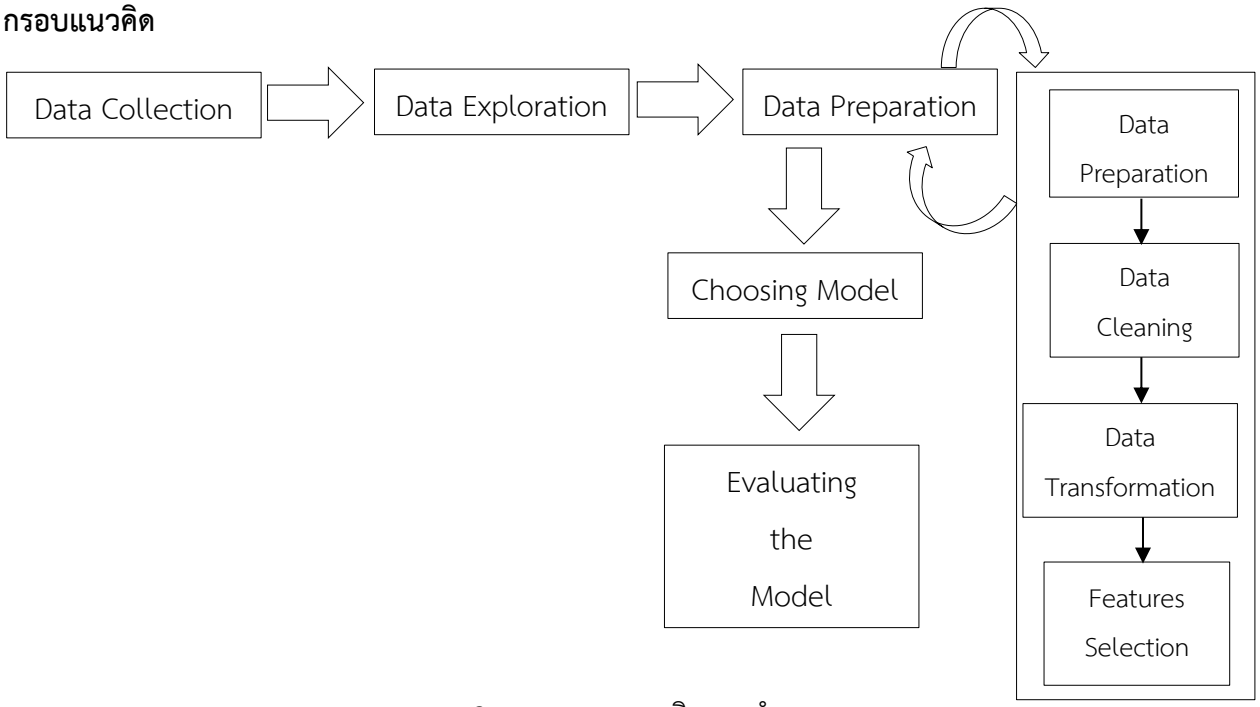
แบ่งข้อมูลออกเป็นชุดข้อมูลสำหรับเรียนรู้ (Training split) 70% และชุดข้อมูลสำหรับทดสอบ (Test split) 30% หลังจากนั้นได้ทำการประเมินประสิทธิภาพอัลกอริทึมทั้งหมดเพื่อเปรียบเทียบประสิทธิภาพในแต่ละอัลกอริทึม ผลการวิจัยพบว่าวิธีต้นไม้ตัดสินใจมีค่าประสิทธิภาพดีที่สุด โดยมีค่าความถูกต้อง (Accuracy) ที่ 100% และค่าความคลาดเคลื่อนกำลังสอง (Mean Squared Error) เฉลี่ยที่ 0.0059

อับดุลเลาะ บากา สุลัยมาน เภอโส๊ะ ฟูไดละห์ ดือมอง และอรรถพล อดุลยศาสน์. (2567) ได้พัฒนาแบบจำลองเพื่อวิเคราะห์ข้อมูลสำหรับการตัดสินใจรับซื้อโคขุน โดยใช้เทคนิคการทำเหมืองข้อมูล 5 วิธี ประกอบด้วย ต้นไม้ตัดสินใจ (Decision Tree) ต้นไม้ตัดสินใจ Iterative Dichotomiser 3 (ID3) นาอีฟเบย (Naïve Bayes) ป่าสุ่ม (Random Forest) และเพื่อนบ้านใกล้สุด (K-Nearest Neighbors) มีการดำเนินการเพื่อจัดการกับข้อมูลที่ผิดพลาด มีการคัดเลือกคุณลักษณะของข้อมูล มีการแบ่งช่วงข้อมูล และใช้ 10-Fold Cross Validation ในการแบ่งข้อมูลออกเป็นชุดข้อมูลสอนและชุดข้อมูลทดสอบ โดยใช้ค่าความถูกต้องในการวัดประสิทธิภาพการพยากรณ์ของแบบจำลอง ผลการทดลองพบว่า เทคนิคการทำเหมืองข้อมูลด้วยวิธีการของป่าสุ่มมีค่าความถูกต้องสูงที่สุดคือ 98.53%

ศรัณยู ชูทองรัตนะ (2565) ได้พัฒนาแบบจำลองพยากรณ์การเกิดโรคไม่ติดต่อเรื้อรังในผู้สูงอายุ ได้แก่ โรคเบาหวาน โรคความดันโลหิตสูง และโรคหัวใจขาดเลือด ด้วยเทคนิคเหมืองข้อมูล ได้แก่ ต้นไม้ตัดสินใจ (Decision Tree), เพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbors), และนาอีฟเบย์ (Naive Bayes) และได้ทำการประเมินประสิทธิภาพแบบจำลอง ผลการวิจัยพบว่า แบบจำลองด้วยเทคนิคต้นไม้ตัดสินใจและเพื่อนบ้านใกล้ที่สุดมีประสิทธิภาพสูงที่สุด โดยมีค่าความถูกต้อง 0.739 ค่าความแม่นยำ 0.728 ค่าความระลึก 0.739

จิราภรณ์ เจริญยิ่ง (2563) ได้พัฒนาแบบจำลองสำหรับจำแนกประเภทข้อมูล การวิเคราะห์องค์ประกอบหลักของข้อมูล เพื่อหาความสัมพันธ์ของตัวแปรและเปรียบเทียบประสิทธิภาพของอัลกอริทึม ซึ่งเป็นการศึกษาด้วยเทคนิคต้นไม้ตัดสินใจ (Decision Tree), ป่าสุ่ม (Random Forest), นาอีฟเบย์ (Naive Bayes) และเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbors) โดยใช้ข้อมูลของนักเรียนระดับมัธยมศึกษาในโรงเรียนโปรตุเกส ผลการวิจัยพบว่า เทคนิคที่ให้ค่าความถูกต้องมากที่สุดคือเทคนิคป่าสุ่ม (Random Forest) เท่ากับ 80.74 และเทคนิคที่มีค่าความถ่วงดุลมากที่สุดคือนาอีฟเบย์ (Naive Bayes) มีค่าเท่ากับ 55.52

**กรอบแนวคิด**




ภาพ 2 แสดงกรอบแนวคิดการทำงาน

จากภาพ 2 อธิบายกรอบแนวคิดการทำงานโดยใช้กระบวนการวิเคราะห์ข้อมูลทางการเรียนรู้ของเครื่องเพื่อใช้กับการพัฒนาโมเดลต้นไม้ตัดสินใจสำหรับการพยากรณ์ข้อมูลด้วยชุดข้อมูลการพยากรณ์โรคหัวใจ เริ่มตั้งแต่ขั้นตอนการเก็บรวบรวมข้อมูล (Data Collection) ถัดมาจะเป็นขั้นตอนการสำรวจข้อมูล (Data Exploration) เพื่อทำความเข้าใจรูปแบบและแนวโน้มของข้อมูล หลังจากนั้นจะเข้าสู่ขั้นตอนการเตรียมข้อมูล (Data Preparation) ซึ่งประกอบด้วยการทำความสะอาดข้อมูล (Data Cleaning) เพื่อลบข้อมูลที่ผิดพลาดหรือขาดหาย การแปลงข้อมูล (Data Transformation) ให้อยู่ในรูปแบบที่เหมาะสมที่สามารถวิเคราะห์ผ่านคอมพิวเตอร์ได้ และการคัดเลือกคุณลักษณะข้อมูล (Features Selection) เพื่อเลือกเฉพาะข้อมูลที่จำเป็นเพื่อการพัฒนาโมเดล จากนั้นจึงทำการเลือกโมเดล (Choosing Model) ที่เหมาะสมสำหรับการวิเคราะห์และการพัฒนาโมเดลเพื่อพยากรณ์ข้อมูล และขั้นตอนสุดท้ายเป็นการประเมินประสิทธิภาพการทำงานของโมเดล (Evaluating the Model) เพื่อวัดประสิทธิภาพและความแม่นยำในการพยากรณ์ข้อมูล

## วิธีดำเนินการวิจัย

### 1. การเก็บรวบรวมข้อมูล (Data Collection)

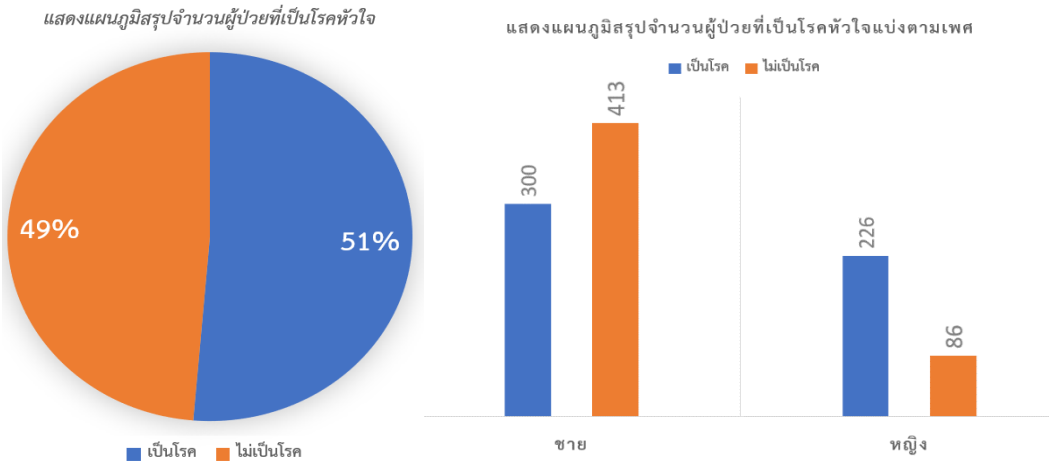
ขั้นตอนนี้ทีมผู้วิจัยได้เก็บรวบรวมข้อมูลจากแหล่งข้อมูลรอง (Secondary Data Sources) โดยใช้ชุดข้อมูล Heart Disease Prediction Dataset จากเว็บไซต์ Kaggle (https://www.kaggle.com/datasets/Johnsmith88/heart-disease-dataset) ซึ่งเป็นชุดข้อมูลที่ใช้ในการพยากรณ์ผู้ป่วยโรคหัวใจ

### 2. การสำรวจข้อมูล (Data Exploration)

ขั้นตอนนี้ทีมผู้วิจัยได้สำรวจข้อมูลจากชุดข้อมูลทั้งหมด 1,025 แถว 14 คอลัมน์ ประกอบด้วยอายุ (Age, เพศ (sex) ประเภทของอาการเจ็บหน้าอก (cp) ความดันโลหิตขณะพัก (trestbps) คอเลสเตอรอล (chol) ระดับน้ำตาลในเลือดขณะอดอาหาร (fbs) คลื่นไฟฟ้าหัวใจขณะพัก (restecg), อัตราการเต้นของหัวใจสูงสุดที่ทำได้ (thalach) อาการเจ็บหน้าอกที่เกิดจากการออกกำลังกาย (exang) ภาวะ ST Depression ที่เกิดจากการออกกำลังกายเมื่อเทียบกับขณะพัก (oldpeak) การวัดมุมหรือลักษณะของการเปลี่ยนแปลงในคลื่น ST ของ ECG ในช่วงการออกกำลังกายที่หนักที่สุด (slope) จำนวนของหลอดเลือดหัวใจหลักที่ได้รับการตรวจด้วยเทคนิค Fluoroscopy (ca), การให้คะแนนความผิดปกติของการไหลเวียนเลือดในกล้ามเนื้อหัวใจ (thal) การเป็นโรคหัวใจ (target) แสดงคอลัมน์และตัวอย่างข้อมูลจากชุดข้อมูล Heart Disease Prediction Dataset ดังภาพ 3

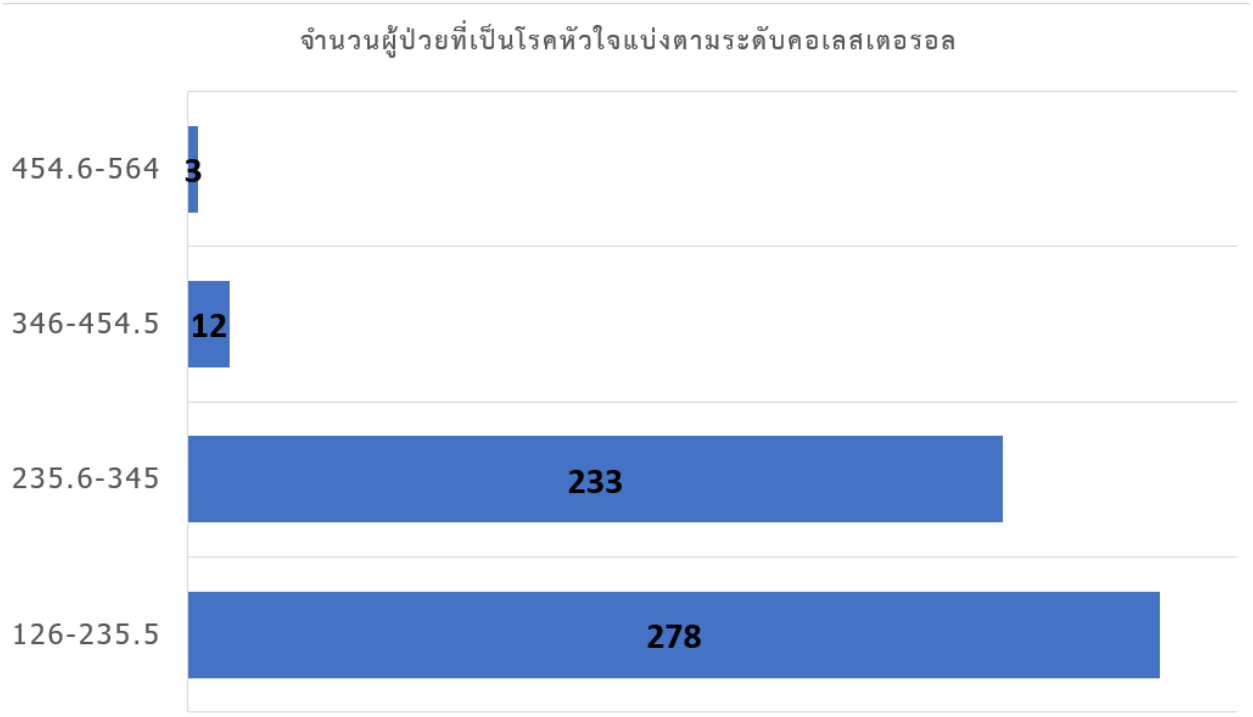
| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | target |
| 2 | 52 | 1 | 0 | 125 | 212 | 0 | 1 | 168 | 0 | 1 | 2 | 2 | 3 | 0 |
| 3 | 53 | 1 | 0 | 140 | 203 | 1 | 0 | 155 | 1 | 3.1 | 0 | 0 | 3 | 0 |
| 4 | 70 | 1 | 0 | 145 | 174 | 0 | 1 | 125 | 1 | 2.6 | 0 | 0 | 3 | 0 |
| 5 | 61 | 1 | 0 | 148 | 203 | 0 | 1 | 161 | 0 | 0 | 2 | 1 | 3 | 0 |
| 6 | 62 | 0 | 0 | 138 | 294 | 1 | 1 | 106 | 0 | 1.9 | 1 | 3 | 2 | 0 |
| 7 | 58 | 0 | 0 | 100 | 248 | 0 | 0 | 122 | 0 | 1 | 1 | 0 | 2 | 1 |
| 8 | 58 | 1 | 0 | 114 | 318 | 0 | 2 | 140 | 0 | 4.4 | 0 | 3 | 1 | 0 |

ภาพ 3 แสดงคอลัมน์และตัวอย่างข้อมูลจากชุดข้อมูล Heart Disease Prediction Dataset




ภาพ 4 แสดงแผนภูมิสรุปจำนวนผู้ป่วยที่เป็นโรคหัวใจ

จากภาพ 4 มีจำนวนผู้ป่วยทั้งหมด 1,025 คน แบ่งเป็นผู้ป่วยที่เป็นโรคหัวใจ 526 คน (51%) และผู้ป่วยที่ไม่เป็นโรคหัวใจ 499 คน (49%) อีกทั้งยังพบว่าเพศชายเป็นโรคหัวใจมากกว่าเพศหญิง โดยที่เพศชายเป็นโรคหัวใจ 300 คน เพศหญิงเป็นโรคหัวใจ 226 คน




ภาพ 5 แสดงแผนภูมิจำนวนผู้ป่วยแบ่งตามระดับคอเลสเตอรอล

จากภาพ 5 จำนวนผู้ป่วยที่เป็นโรคหัวใจแบ่งตามระดับคอเลสเตอรอล โดยค่าคอเลสเตอรอล 454.6 -564 มีจำนวน 3 คน, 346 – 454.5 มีจำนวน 12 คน, 235.6 -345 มีจำนวน 233 คน, และ 126 – 235.5 มีจำนวนมากที่สุดที่ 278 คน

**3. การเตรียมข้อมูล (Data Preparation)**

**3.1 การทำความสะอาดข้อมูล (Data Cleaning)**

ในขั้นตอนนี้ ทีมผู้วิจัยได้ทำการตรวจสอบข้อมูลที่ผิดปกติ พบว่าไม่มีข้อมูลสูญหาย (Missing values) และไม่มีข้อมูลซ้ำซ้อน (Duplicates) จึงสามารถนำชุดข้อมูลนี้ไปใช้งานได้โดยไม่จำเป็นต้องดำเนินการทำความสะอาดข้อมูลเพิ่มเติม

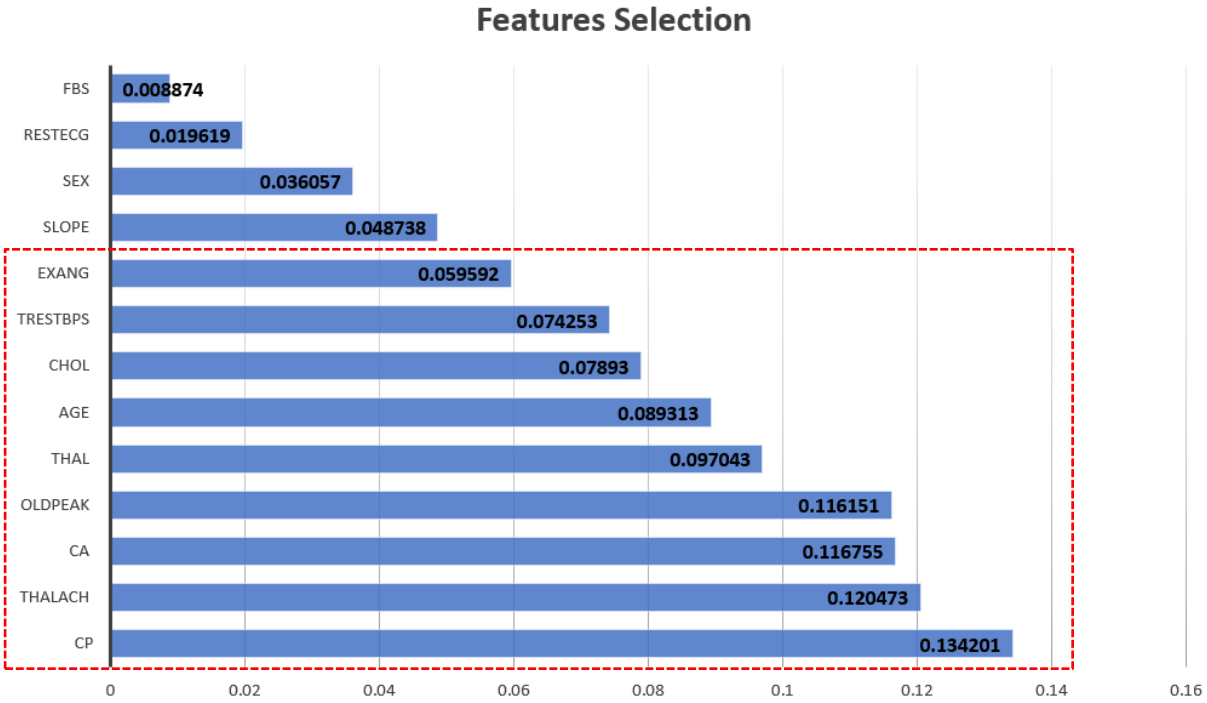
**3.2 การแปลงข้อมูล (Data Transformation)**

ในขั้นตอนนี้ ทีมผู้วิจัยได้ดำเนินการตรวจสอบชนิดของข้อมูลในทุกคอลัมน์ พบว่าเป็นข้อมูลในรูปแบบตัวเลข ซึ่งถือเป็นข้อมูลที่พร้อมสำหรับการนำไปวิเคราะห์และสร้างโมเดล ดังนั้นจึงไม่จำเป็นต้องดำเนินการแปลงข้อมูลเพิ่มเติม

**3.3 การคัดเลือกคุณลักษณะข้อมูล (Features Selection)**

ในขั้นตอนนี้ ทีมผู้วิจัยได้ทำการเลือกคุณลักษณะข้อมูลด้วยเทคนิคป่าสุ่ม (Random Forest) โดยนำข้อมูลทุกคอลัมน์จากชุดข้อมูล เพื่อวิเคราะห์หาคุณลักษณะที่สำคัญ (Features importance) โดยใช้

ความสามารถของ RandomForestClassifier ในการประเมินความสำคัญของแต่ละคุณลักษณะ แสดงค่าคุณลักษณะข้อมูลที่สำคัญดังภาพ 6




ภาพ 6 แสดงค่าคุณลักษณะข้อมูลที่สำคัญ

จากภาพ 6 แสดงค่าคุณลักษณะข้อมูลที่สำคัญทั้งหมด 13 คอลัมน์ และได้ทำการคัดเลือกค่าคุณลักษณะที่มากกว่า 0.05 ทั้งหมด 9 คอลัมน์ ประกอบด้วย คอลัมน์ CP (0.134201), THALACH (0.120473), CA (0.116755), OLDPEAK (0.116151), THAL (0.097043), AGE (0.089313), CHOL (0.07893), TRESTBPS (0.074253), และ EXANG (0.059529) มาใช้สำหรับพัฒนาโมเดล

**4. การเลือกโมเดล (Choosing Model)**

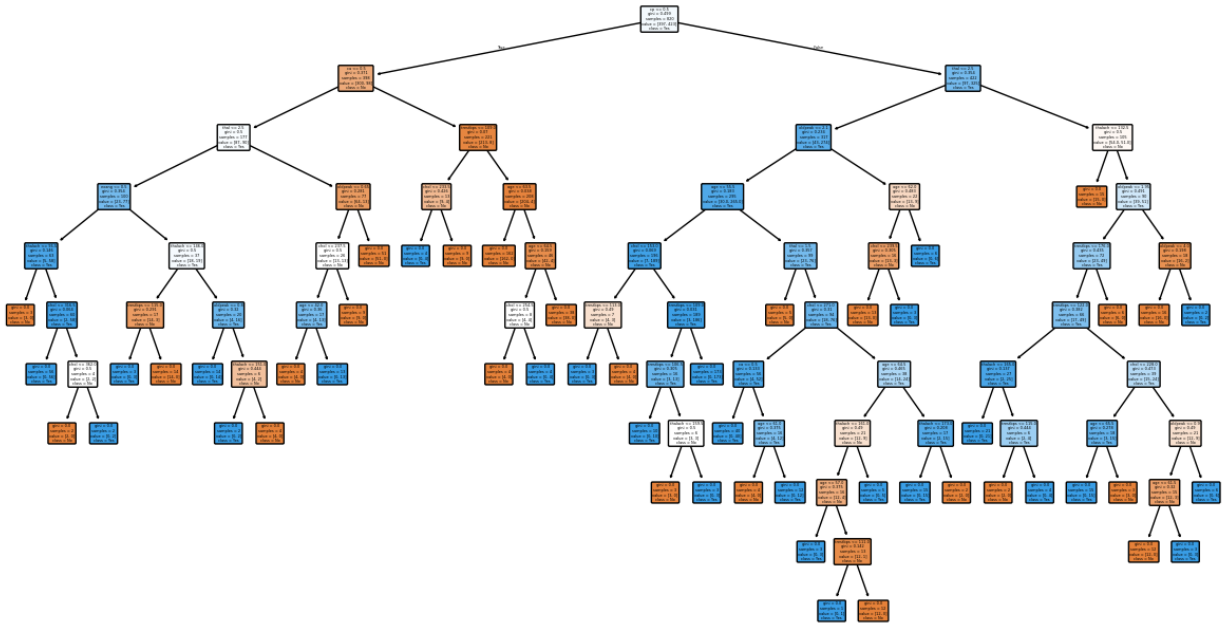
ในขั้นตอนนี้ทีมผู้วิจัยเลือกโมเดลต้นไม้ตัดสินใจ (Decision tree) ในการพัฒนาโมเดลเพื่อพยากรณ์ผู้ป่วยโรคหัวใจ เนื่องจากเป็นโมเดลที่มีความแม่นยำและได้รับความนิยมเกี่ยวกับการวิเคราะห์ข้อมูลทางการแพทย์ ได้ทำการแบ่งชุดข้อมูลสำหรับเรียนรู้ (Training set) 80% และชุดข้อมูลสำหรับทดสอบ (Test set) 20%

**5. การประเมินประสิทธิภาพการทำงานของโมเดล (Evaluating the Model)**

หลังจากพัฒนาโมเดลเสร็จแล้ว ในขั้นตอนนี้ทีมผู้วิจัยทำการประเมินผลโมเดลด้วย 4 ตัวชี้วัด ประกอบด้วยค่าความถูกต้อง (Accuracy), ความแม่นยำ (Precision), ความระลึก (Recall), และคะแนน F1 (F1-score)

## ผลการวิจัย

งานวิจัยชิ้นนี้ได้ทำการพัฒนาโมเดลต้นไม้ตัดสินใจสำหรับการพยากรณ์ข้อมูลด้วยชุดข้อมูลการพยากรณ์โรคหัวใจ โดยเก็บรวบรวมข้อมูลจากแหล่งข้อมูลรอง ใช้ชุดข้อมูล Heart Disease Prediction Dataset จากเว็บไซต์ Kaggle ซึ่งเป็นชุดข้อมูลที่ใช้ในการพยากรณ์ผู้ป่วยโรคหัวใจจำนวน 1,025 แถว 14 คอลัมน์ ได้ใช้กระบวนการการเรียนรู้ของเครื่องเพื่อดำเนินการวิจัย เริ่มตั้งแต่ขั้นตอนการเก็บรวบรวมข้อมูล การสำรวจข้อมูล การเตรียมข้อมูล การเลือกโมเดล ตลอดจนกระทั่งการประเมินประสิทธิภาพการทำงานของโมเดล แสดงโครงสร้างต้นไม้ตัดสินใจดังภาพ 7



ภาพ 7 แสดงโครงสร้างต้นไม้ตัดสินใจ

การประเมินประสิทธิภาพการทำงานของโมเดล ผู้วิจัยใช้โปรแกรมภาษา Python ในการประเมินประสิทธิภาพโดยใช้ 4 ตัวชี้วัด ประกอบด้วยค่าความถูกต้อง (Accuracy) ความแม่นยำ (Precision) ความระลึก (Recall) และคะแนน F1 (F1-score) แสดงผลการประเมินประสิทธิภาพการทำงานของโมเดลดังตาราง 1

ตาราง 1 แสดงผลการประเมินประสิทธิภาพการทำงานของโมเดล

| **ตัวชี้วัด** | **ผลการประเมิน** | |
|---|---|---|
| | **ค่าที่ได้** | **%** |
| ค่าความถูกต้อง (Accuracy) | 0.9854 | 98.54 |
| ค่าความแม่นยำ (Precision) | 1.0000 | 100 |
| ค่าความระลึก (Recall) | 0.9709 | 97.09 |
| ค่าคะแนน F1 (F1-score) | 0.9852 | 98.52 |

## สรุปและอภิปรายผลการวิจัย

ในงานวิจัยนี้ ทีมผู้วิจัยได้ดำเนินการพัฒนาโมเดลการพยากรณ์โรคหัวใจโดยใช้ชุดข้อมูล Heart Disease Prediction Dataset ที่มีข้อมูลทั้งหมด 1,025 แถวและ 14 คอลัมน์ ซึ่งรวมถึงคุณลักษณะต่าง ๆ เช่น อายุ (age) เพศ (sex), ความดันโลหิต (trestbps), คอเลสเตอรอล (chol) และอื่น ๆ โดยใช้ชุดข้อมูลจากแหล่งข้อมูลรอง (Secondary Data Sources) จากเว็บไซต์ Kaggle ในขั้นตอนการสำรวจข้อมูล ทีมผู้วิจัยได้พบว่า ชุดข้อมูลนี้มีผู้ป่วยโรคหัวใจ 526 คน (51%) และไม่เป็นโรคหัวใจ 499 คน (49%) โดยมีข้อมูลเกี่ยวกับคอเลสเตอรอลที่ถูกแบ่งเป็นหลายช่วงและพบว่าเพศชายมีจำนวนผู้ป่วยโรคหัวใจมากกว่าเพศหญิง สำหรับการเตรียมข้อมูล ทีมผู้วิจัยได้ทำการตรวจสอบข้อมูลสูญหายและข้อมูลซ้ำซ้อน ซึ่งพบว่าไม่มีข้อมูลสูญหายและข้อมูลซ้ำซ้อน จึงไม่จำเป็นต้องดำเนินการทำความสะอาดข้อมูล นอกจากนี้ยังได้ตรวจสอบชนิดของข้อมูลและพบว่าเป็นข้อมูลตัวเลขที่พร้อมสำหรับการนำไปวิเคราะห์ ในการเลือกคุณลักษณะข้อมูล (Feature Selection) ทีมผู้วิจัยใช้เทคนิคป่าสุ่ม (Random Forest) เพื่อคัดเลือกคุณลักษณะที่มีความสำคัญมากกว่า 0.05 ซึ่งได้ผลลัพธ์ว่ามีคอลัมน์ที่มีความสำคัญ 9 คอลัมน์ โดยมีคอลัมน์ CP, THALACH, CA, OLDPEAK, THAL, AGE, CHOL, TRESTBPS, และ EXANG เป็นคุณลักษณะที่สำคัญต่อการพยากรณ์โรคหัวใจ ในการเลือกโมเดล ทีมผู้วิจัยเลือกใช้ต้นไม้ตัดสินใจ (Decision Tree) เนื่องจากเป็นโมเดลที่เข้าใจง่ายและสามารถตีความผลลัพธ์ได้ชัดเจน จากนั้นทำการฝึกโมเดลโดยใช้ข้อมูล 80% สำหรับการฝึกและ 20% สำหรับการทดสอบ ผลการประเมินโมเดลที่ได้จากการทดสอบด้วยตัวชี้วัดต่าง ๆ ได้แก่ Accuracy, Precision, Recall, และ F1-score แสดงให้เห็นว่าโมเดลที่พัฒนาออกมามีประสิทธิภาพสูง โดยได้ผลลัพธ์ Accuracy เท่ากับ 98.54%, Precision เท่ากับ 100%, Recall เท่ากับ 97.09% และ F1-score เท่ากับ 98.52% ผลการวิจัยนี้แสดงให้เห็นว่าโมเดลต้นไม้ตัดสินใจสามารถพยากรณ์โรคหัวใจได้อย่างมีประสิทธิภาพและแม่นยำ โดยมีค่าตัวชี้วัดที่แสดงถึงความสามารถในการทำนายที่สูงมาก ซึ่งอาจนำไปสู่การพัฒนาเครื่องมือหรือแอปพลิเคชันที่ช่วยในการคัดกรองผู้ป่วยโรคหัวใจในอนาคตได้อย่างมีประสิทธิภาพ

ผลการวิจัยนี้มีความสำคัญอย่างยิ่งในแง่ของการพัฒนาเครื่องมือหรือแอปพลิเคชันสำหรับการคัดกรองผู้ป่วยโรคหัวใจโดยการใช้โมเดลต้นไม้ตัดสินใจที่มีประสิทธิภาพสูงและการประเมินผลที่ดี การนำโมเดลนี้ไปใช้งานในทางการแพทย์สามารถช่วยให้การตรวจจับโรคหัวใจทำได้อย่างรวดเร็วและแม่นยำ ซึ่งจะเป็นประโยชน์ในการช่วยลดการเกิดโรคหัวใจในระยะยาว อย่างไรก็ตาม ผลการศึกษานี้ได้จากการใช้ชุดข้อมูล Heart Disease Prediction Dataset จาก Kaggle ซึ่งอาจจะมีข้อจำกัดในแง่ของความหลากหลายของข้อมูลที่ใช้ เนื่องจากข้อมูลที่ใช้ในงานวิจัยนี้มีจำนวน 1,025 แถว อาจจะไม่ครอบคลุมผู้ป่วยโรคหัวใจจากหลากหลายปัจจัยหรือภูมิภาค ดังนั้น การทดสอบโมเดลด้วยชุดข้อมูลที่หลากหลายและข้อมูลใหม่ในอนาคตจึงเป็นสิ่งสำคัญในการยืนยันความสามารถของโมเดล

งานวิจัยนี้แสดงให้เห็นว่า โมเดลต้นไม้ตัดสินใจ สามารถพยากรณ์โรคหัวใจได้อย่างมีประสิทธิภาพและแม่นยำ โดยใช้ข้อมูลจาก Heart Disease Prediction Dataset และใช้เทคนิคต่าง ๆ เช่น Random Forest ใน

การเลือกคุณลักษณะที่สำคัญ ผลการประเมินจากตัวชี้วัดต่าง ๆ ได้แสดงให้เห็นถึงความสำเร็จในการสร้างโมเดลที่มีความแม่นยำสูง ซึ่งสามารถนำไปพัฒนาเครื่องมือทางการแพทย์ที่ช่วยในการคัดกรองผู้ป่วยโรคหัวใจในอนาคตได้

## เอกสารอ้างอิง

วุรวัชร ศรีเปารยะ และและสายชล สินสมบูรณ์ทอง. (2560). การเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มการเป็นโรคไตเรื้อรัง: กรณีศึกษาโรงพยาบาลแห่งหนึ่งในประเทศอินเดีย. *วารสารวิทยาศาสตร์และเทคโนโลยี*, 25(5), 839-853.

อับดุลเลาะ บากา, สุลัยมาน เภอโส๊ะ, ฟูไดละห์ ดือมอง และอรรถพล อดุลยศาสน์. (2567). แบบจำลองเพื่อการตัดสินใจรับซื้อโคขุน โดยใช้เทคนิคการทำเหมืองข้อมูล กรณีศึกษา กลุ่มวิสาหกิจชุมชนโคเนื้อบ้านบูเกะจือฆา. *วารสารแม่โจ้เทคโนโลยีสารสนเทศและนวัตกรรม*, 10(1), 99-117.

ศรัณยู ชูทองรัตนะ. (2565). *แบบจำลองเพื่อพยากรณ์การเกิดโรคไม่ติดต่อเรื้อรังในผู้สูงอายุ ด้วยเทคนิคเหมืองข้อมูล: กรณีศึกษาโรงพยาบาลนาหว้า จังหวัดนครพนม* [วิทยานิพนธ์ปริญญามหาบัณฑิต ที่ไม่มีการตีพิมพ์]. มหาวิทยาลัยราชภัฏสกลนคร.

จิราภรณ์ เจริญยิ่ง. (2563). *การพยากรณผลสัมฤทธิ์ทางการเรียนด้วยเทคนิคเหมืองข้อมูลโดยใช้ Rapid Miner* [สารนิพนธ์วิทยาศาสตรมหาบัณฑิต ที่ไม่มีการตีพิมพ์]. มหาวิทยาลัยศรีนครินทรวิโรฒ.

ศุภจิตรี บุญอำนวย. (2562). *การปรับปรุงประสิทธิภาพต้นไม้ตัดสินใจแบบจำแนกและแบบถดถอยด้วยเทคนิคการสุ่มซ้ำ* [วิทยานิพนธ์วิศวกรรมศาสตรมหาบัณฑิต ที่ไม่มีการตีพิมพ์]. มหาวิทยาลัยเทคโนโลยีสุรนารี.