การประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลสำหรับการจำแนกประเภทข้อมูลเห็ด
Application of Data Mining Techniques for Mushroom Data Classification

ชัยนคร สาริสี [1] และ นิภาพร ชนะมาร [2]
chainakon sarisee [1]and Nipaporn Chanamarn [2]
สาขาวิชาคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏสกลนคร 47000
Department of Computer, Faculty of Science and technology, Sakon Nakhon Rajabhat University, Sakon Nakhon Province, 47000
Corresponding author Email: nipaporn@snru.ac.th

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของการจำแนกข้อมูลด้วยอัลกอริทึมเหมืองข้อมูลสามแบบคือโครงข่ายประถาทเทียม การเรียนรู้แบบเบย์และต้นไม้ตัดสินใจ เพื่อให้ได้อัลกอริทึมที่มีประสิทธิภาพสูงสุดที่จะถูกนำมาวิเคราะห์ข้อมูลเห็ด จากปัญหาเห็ดพิษ ซึ่งข้อมูลที่นำมาใช้ ในการทดลองเป็นข้อมูลจาก ฐานข้อมูล UCI Mushroom จาก ปี1987 ซึ่งมีจำนวนข้อมูล 8,124 ชุดข้อมูล และปัจจัยสำหรับการวิเคราะห์ 22 ปัจจัย โดยเปรียบเทียบประสิทธิภาพของแบบจำลองด้วย โปรแกรม Weka 3.8 ผลการวิจัยพบว่า การจำแนกประเภทข้อมูลด้วย ต้นไม้ตัดสินใจ (Decision Tree) กับโครงข่ายประสาทเทียม (Artificial Neural network) มีประสิทธิภาพเท่ากัน โดยมีค่าความถูกต้อง 100% ค่าความแม่นยำ 100% ค่าความระลึก 100% ค่าความถ่วงดุล 100% และ การเรียนรู้แบบเบย์(Naive Bayes) มีค่าต่ำสุดของทั้ง3อัลกอริทึม โดยมีค่าความถูกต้อง 92.40% ค่าความแม่นยำ 95.40% ค่าความระลึก 99.30% ค่าความถ่วงดุล 95.70%

คำสำคัญ: เห็ด, ต้นไม้ตัดสินใจ, การเรียนรู้แบบเบย์, โครงข่ายประสาทเทียม, การเรียนรู้ของเครื่องแบบมีผู้สอน

ABSTRACT

The objective of this research is to compare the performance of data classification using three different data mining algorithms, namely Artificial Neural Networks, Naive Bayes, and Decision Trees, in order to identify the most efficient algorithm for analyzing mushroom data, particularly data related to poisonous in the mushrooms. The data used for experimentation is sourced from the UCI Mushroom database from 1987, which comprises 8,124 data sets and 22 factors for analysis. The performance of these models was compared using Weka 3.8 software. The research findings revealed that classifying data using Decision Trees and Artificial Neural Networks yielded identical performance, with an accuracy, precision, recall, and F-measure. all reaching 100%. However, Naive Bayes exhibited the lowest performance among the three algorithms, with an accuracy of 92.40%, precision of 95.40%, recall of 99.30%, and F-measure.of 95.70%.

Keywords: Mushroom, Decision Tree, Naive baye, Neural Network and Supervised Learning

บทนำ

เห็ดเป็นอาหารที่มีความนิยมและมีความหลากหลายมากในทุกส่วนของโลก การระบุและการจำแนกเห็ดเป็นกลุ่มที่ปลอดภัยและไม่ปลอดภัยมีความสำคัญอย่างมากเพื่อป้องกันอันตรายต่อสุขภาพของผู้บริโภค หนึ่งในปัญหาหลักของเห็ดคือความซับซ้อนในการระบุชนิดของเห็ด มีเห็ดที่ปลอดภัยทานได้และมีคุณค่าทางโภชนาการ แต่มีเห็ดอื่นๆ ที่อาจเป็นพิษหรือไม่เหมาะสมทาน ปัญหาที่สำคัญเมื่อพบผู้ป่วยจากการรับประทานเห็ดพิษ(ดร.ปริญญา จันทรศรี,2563) ส่วนใหญ่แล้วแพทย์หรือแม้แต่ผู้ป่วยไม่รู้จักเห็ดชนิดนั้น อย่างไรก็ตาม ในเห็ดพิษชนิดเดียวกันอาจมีสารพิษอยู่หลายชนิดต่างๆ กัน ตามพื้นที่ที่เห็ดขึ้น รวมทั้งการพิสูจน์ว่าเป็นเห็ดชนิดใด อาจต้องใช้เวลามากจนให้การรักษาไม่ทันการ ซึ่งในประเทศไทยในหน้าฝนมักจะเจอกับเห็ดระโงกหิน เห็ดระโงกดำพิษ เห็ดระงากขาว และ เห็ดไข่ตายซาก หลังแสดงอาการประมาณ 1 วัน อาจมีอาการตับอักเสบรุนแรงถึงขั้นเสียชีวิต (โรงพยาบาลจุฬาลงกรณ์ สภากาชาดไทย,อ. ดร. พญ.กรวลี มีศิลปวิกภัย,2564)

ปัจจุบันเทคโนโลยีสมัยใหม่มีการประยุกต์ใช้อย่างมากมาย โดยเฉพาะการใช้เทคโนโลยีการเรียนรู้ของเครื่องและการทำเหมืองข้อมูล ซึ่งเป็นเทคนิคที่ท้าทายและที่มีศักยภาพมากในการแก้ไขปัญหาในหลากหลายด้าน เราได้เห็นการนำเทคโนโลยีนี้ไปประยุกต์ใช้ในหลายกลุ่มงาน เช่นด้านการแพทย์ การศึกษา การเกษตร และด้านสิ่งแวดล้อม งานวิจัยในหลายด้านก็ได้เริ่มต้นเปิดโอกาสในการนำเทคโนโลยีเข้ามาช่วยในการแก้ไขปัญหาของโลกในรูปแบบที่มิได้คาดคิดมาก่อน ในบทความนี้จะสรุปความสำคัญของการวัดประสิทธิภาพของโมเดลการจำแนกประเภท ที่มีความสำคัญอย่างยิ่งเมื่อนำไปใช้ในสถานการณ์จริง เข้าใจหลักการและสิ่งสำคัญที่เกี่ยวข้องในเรื่องนี้ โดยมีงานวิจัยของ (วทัญญู ชูประจิตต์, ผู้ช่วยศาสตราจารย์ดร.นพมาศ ปักเข็ม และ ผู้ช่วยศาสตราจารย์ดร.สิรยา สิทธิสาร) ทำการวิจัยเรื่องการจำแนกประเภทและวิธีการคัดเลือกคุณลักษณะสำหรับการสนับสนุนการวิเคราะห์ความรู้สึกต่อสถานที่ท่องเที่ยวไทย การวิเคราะห์ความรู้สึกต่อสถานที่ท่องเที่ยวไทยจากข้อมูลการแสดงความคิดเห็นต่อสถานที่ท่องเที่ยวไทยที่เก็บรวบรวมจากสื่อสังคมออนไลน์ภายในประเทศไทย การจำแนกประเภทใช้กระบวนการคัดเลือกคุณลักษณะแบบหาความถี่คำและสร้างตัวจำแนกประเภทด้วย 4 อัลกอริทึม คือ ขั้นตอนวิธีการค้นหาเพื่อนบ้านใกล้สุด เค ตัว,ต้นไม้ตัดสินใจ, นาอีฟเบย์ และ ซัพพอร์ตเวกเตอร์แมชชีน ด้วยโปรแกรม WEKA เพื่อทดสอบความแม่นยำของโมเดล การทดสอบความถูกต้องจากผลทดสอบพบว่าอัลกอริทึมมซัพพอร์ตเวกเตอร์แมชชีน มีค่าความถูกต้องสูงสุด 99.33 % ส่วนงานวิจัยของ(เจษฎาพร ปาคำวัง, กาญจน คุ้มทรัพย์ และ วรชัย ศรีเมือง) การเปรียบเทียบแบบจำลองจำแนกประเภทเพื่อการพัฒนาการท่องเที่ยว อำเภอเขาค้อ จังหวัดเพชรบูรณ์ด้วยเทคนิคเหมืองข้อมูล โดยใช้อัลกอริทึม Bayes network classifier,Naive bayes,Logistic regression,Sequential minimal optimization,Decision tree และ Random forest ด้วยโปรแกรม RapidMiner Studio เพื่อทดสอบความแม่นยำของโมเดล จากการทดสอบประสิทธิภาพอัลกอริทึมทั้ง 7 อัลกอริทึม พบว่า Decision tree ให้ค่าความถูกต้องสูงที่สุดอยู่ที่ 94.57% และงานวิจัยของณัฐวดี หงส์บุญมี และ ประภาสิริ ตรีพาณิชกุล การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลเพื่อวิเคราะห์ปัจจัยความเสี่ยงที่ส่งผลต่อการเกิดโรคไฮเปอร์ไทรอยด์ด้วยเทคนิคเหมืองข้อมูลโดยใช้อัลกอริทึมโครงข่ายประสาทเทียม,การเรียนรู้แบบเบย์และต้นไม้ตัดสินใจด้วยโปรแกรม WEKA 3.8 เพื่อทดสอบความแม่นยำของโมเดล จากการทดสอบประสิทธิภาพอัลกอริทึมทั้ง 3 อัลกอริทึมม พบว่าอัลกอริทึมโครงข่ายประสาทเทียมให้ค่าความถูกต้องสูงที่สุดอยู่ที่ 82.97%

การประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลในการประเมินและประเภทข้อมูลเห็ดอาจช่วยให้เรามีข้อมูลที่มีประโยชน์เกี่ยวกับเห็ดในหลายด้าน เช่น ความปลอดภัย การตรวจสอบความสมบูรณ์ทางโภชนาการ หรือแนวโน้มในการเจริญเติบโตของเห็ด โดยมีการเปรียบประสิทธิภาพแบบจำลองเพื่อหาแบบจำลองที่เหมาะสมที่สุด จากอัลกอริทึม 3 เทคนิค คือ Decision Tree, Naive baye และ Neural Network โดยใช้ชุดข้อมูล Mushroom จากฐานข้อมูล UCI ปี1987 ซึ่งมีจำนวนข้อมูล 8,124 ชุดข้อมูล และปัจจัยสำหรับการวิเคราะห์ 22 ปัจจัย

เอกสารที่เกี่ยวข้อง

1. การเหมืองข้อมูล

การทำเหมืองข้อมูล (Data Mining) คือ การค้นหาข้อมูลที่มีประโยชน์จากแหล่งข้อมูลที่มีเป็นจำนวนมากมายมหาศาล เพื่อดึงข้อมูลที่มีประโยชน์มาทำการวิเคราะห์ค้นหารูปแบบหรือความสัมพันธ์ที่เกิดในฐานข้อมูล และจัดทำเป็นสารสนเทศเพื่อใช้ในการวางแผนบริหารจัดการธุรกิจ โดยการแยกข้อมูลที่มีประโยชน์ออกมาใช้งานเปรียบเทียบคล้ายกับการทำเหมืองแร่ ที่จะต้องทำการแยกเศษหินดินทรายที่ไม่มีค่าและมีปริมาณมากออกจากแร่ที่มีมูลค่ามากและมักจะมีปริมาณน้อย ในการวิจัยนี้ใช้อัลกอริทึมการจำแนกประเภทข้อมูล(Classification) เป็นเทคนิคการวิเคราะห์จากการเรียนรู้จดจำหรือจากรูปแบบของข้อมูลแบบมีป้ายบอกทาง ซึ่งเป็นการเรียนรู้ของเครื่องแบบมีผู้สอน (Supervised Machine Learning Techniques) (rattanatat,2019)

2. การเรียนรู้ของเครื่องแบบมีผู้สอน

การเรียนรู้ของเครื่องแบบมีผู้สอน (Supervised Learning) เป็นศาสตร์แขนงหนึ่งใน AI หรือปัญญาประดิษฐ์ ภายใต้หัวข้อ Machine Learning ที่กำลังเป็นที่นิยมในการศึกษาและวิจัยกันในปัจจุบัน เนื่องจากทำได้ง่าย ต้นทุนต่ำ เพียงใช้คอมพิวเตอร์เครื่องเดียวก็สามารถศึกษาและทำงานจนเห็นผลได้เลย แล้วศาสตร์แขนงนี้มีมานานมากแล้วตั้งแต่ปี 1959 ถูกเสนอโดย Arthur Samuel เป็นนักวิทยาศาสตร์คอมพิวเตอร์ชาวอเมริกันผู้เชี่ยวชาญด้านเกมคอมพิวเตอร์ ปัญญาประดิษฐ์และการเรียนรู้ของเครื่อง โปรแกรม แต่ด้วยเทคโนโลยีหรือระบบประมวลผลในตอนนั้นยังล้าสมัยอยู่ ทำให้ยังไม่เป็นที่นิยม ผิดกับในปัจจุบัน (Natdanai James,2019)

3. เทคนิคต้นไม้ตัดสินใจ

เทคนิคต้นไม้ตัดสินใจ (Decision Tree) คือ แบบจำลองทางคณิตศาสตร์ เพื่อการหาทางเลือกที่ดีที่สุด โดยการนำข้อมูลมาสร้างแบบจำลองการพยากรณ์ในรูปแบบของ โกร งสร้างต้นไม้ ซึ่งมีการเรียนรู้ข้อมูลแบบมีผู้สอน (Supervised Leaming) สามารถสร้างแบบจำลองการจัดหมวดหมู่(Clustering) ได้จากกลุ่มตัวอย่างของข้อมูลที่กำหนดไว้ล่วงหน้า (Training scl) ได้โดยอัตโนมัติและสามารถพยากรณ์กชุ่มของรายการที่ยังไม่เคยนำมาจัดหมวดหมู่ได้อีกด้วยโดยปกติมักประกอบด้วยกฎในรูปแบบ "ถ้า เงื่อนไข แล้ว ผลลัพธ์" (รุจิรา ธรรมสมบัติ, 2554)

เช่น "If Income = High and Married = No THEN Risk = Poor"

"If Income = High and Married = Yes THEN Risk = Good"

4. เทคนิคการเรียนรู้แบบเบย์

Naive Bayes Classification เป็นหนึ่งใน Classification Model ใช้ในการแบ่งกลุ่มหรือหาเหตุการณ์ที่จะเกิดขึ้นโดยการอิงทฤษฎีความน่าจะเป็นของ Bayes หรือ Bayesian ซึ่ง Target ของโมเดลจะมีความคล้ายคลึงกับ Logistic Regression ว่าจะเกิดเหตุการณ์นั้นหรือไม่โดยจะเพิ่มโอกาสในการเกิดเหตุการณ์เข้าไปด้วย โดยมักจะใช้ในการวิเคราะห์ข้อมูลที่มีความต่อเนื่องของเหตุการณ์ (Dependent Event) เช่น โอกาสในการเกิดโรคในกลุ่มประชากรที่เราสนใจ ซึ่งจำเป็นจะต้องอาศัยการคำนวณผ่านสูตรดังนี้ และกำหนดให้(สถาบันนวัตกรรมและธรรมาภิบาลข้อมูล,2565)

$$P(A|B)= \frac{P(B|A)\ P(A)}{P(B)}$$

p(X) คือความน่าจะเป็นในการเกิดเหตุการณ์ X

1 – p(X) คือความน่าจะเป็นที่จะไม่เกิดเหตุการณ์ X

exp คือการหา Exponential ของสมการ

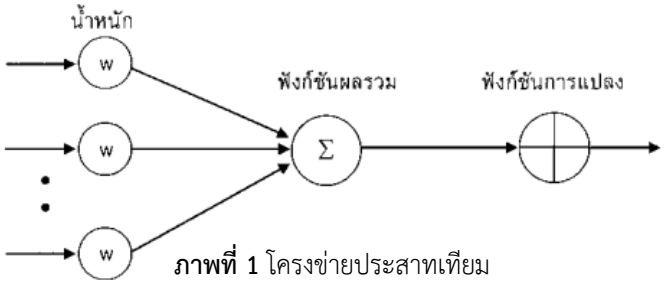
B0 คือค่า Intercept หรือค่าคงที่ซึ่งจะเป็นค่าที่ส่งผลต่อสมการ Regression

B1 คือค่า Parameter หรือค่าสัมประสิทธิ์ของตัวแปรอิสระ

X ตัวแปรอิสระ

6. เทคนิคโครงข่ายประสาทเทียม

โครงข่ายประสาทเทียม (Artificial Neural network) เป็นศาสตร์แขนงหนึ่งทางด้านปัญญาประดิษฐ์(Artificial Intelligence : AI) (ธนาวุฒิ ประกอบผล, 2552) ที่สามารถนำไปประยุกต์ใช้กับงานหลายด้านได้อย่างมีประสิทธิภาพ เช่นการจำแนกรูปแบบ การทำนาย การควบคุม การหาความเหมาะสม และการจัดกลุ่ม เป็นต้นหลักการสำคัญของโครงข่ายประสาทเทียม คือ ความพยายามที่จะลอกเลียนแบบการทำงานของเซลล์ประสาทในสมองมนุษย์เพื่อทำงานได้อย่างมี ประสิทธิภาพ ลักษณะทั่วไปของโครงข่ายประสาทเทียม คือ การที่โหนด (node) ต่าง ๆ จำลองมาจากไซแนป (synapse) ของเซลล์ประสาทระหว่างเดนไดร์ท (dendrite) และแอกซอน (axon) โดยมีฟังก์ชันเป็นตัวกำหนดสัญญาณส่งออก (activationfunction or transfer function) นั่นเองลักษณะของโครงข่ายประสาทเทียมสามารถแบ่งได้ 2 แบบ คือ 1) โครงข่ายประสาทเทียมแบบชั้นเดียว (single layer)ซึ่งจะมีเพียงชั้นสัญญาณประสาทขาเข้า และชั้นสัญญาณประสาทขาออกเท่านั้น เช่น โครงข่ายเพอเซบตรอนอย่างง่าย (simple perceptron) และโครงข่ายโฮปฟิลด์ (hopfieldnetworks) เป็นต้น และ 2) โครงข่ายประสาทเทียมแบบหลายชั้น (multi layer) ซึ่งมีลักษณะเช่นเดียวกับโครงข่ายประสาทเทียมแบบชั้นเดียว แต่จะมีชั้นแ อบแฝง (hidden) เพิ่มขึ้น โดยอยู่ส่วนกลางระหว่างชั้นนำข้อมูลป้อนเข้าและชั้นส่งข้อมูลออก



ภาพที่ 1 โครงข่ายประสาทเทียม

7. งานวิจัยที่เกี่ยวข้อง

ผู้วิจัยได้ทำการค้นคว้างานวิจัยด้านเหมืองข้อมูลที่เกี่ยวข้องวิธีการนำมาใช้ในการวิจัย เพื่อนำมาเป็นข้อมูลในการศึกษาและอ้างอิง ผู้วิจัยได้ศึกษางานวิจัยที่มีความเกี่ยวข้อง ดังต่อไปนี้

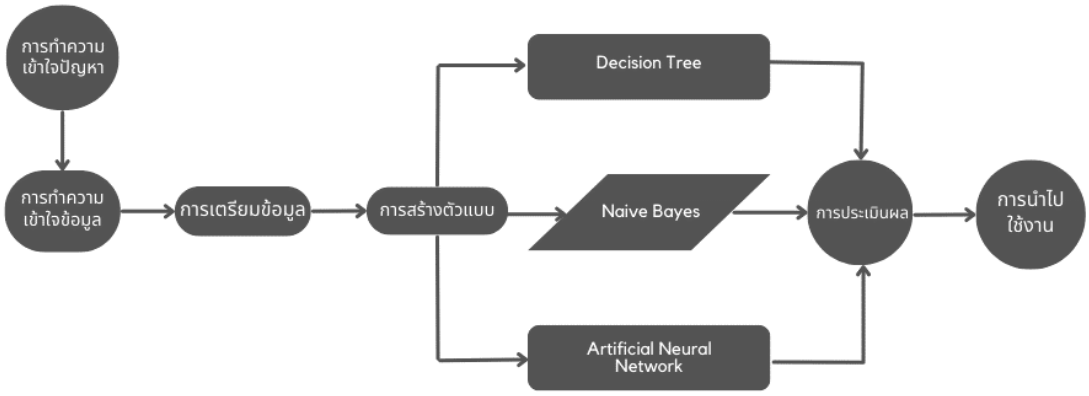
วทัญญู ชูประจิตต์, ผู้ช่วยศาสตราจารย์ดร.นพมาศ ปักเข็ม และ ผู้ช่วยศาสตราจารย์ดร.สิรยา สิทธิสาร การจำแนกประเภทและวิธีการคัดเลือกคุณลักษณะสำหรับการสนับสนุนการวิเคราะห์ความรู้สึกต่อสถานที่ท่องเที่ยวไทย โดยใช้อัลกอริทึมต้นไม้ตัดสินใจ,วิธีการเพื่อนบ้านใกล้ที่สุด, เนอีฟเบย์และเทคนิคซัพพอร์ตเวกเตอร์แมชชีน จากการทดสอบประสิทธิภาพอัลกอริทึมทั้ง 4 อัลกอริทึมม พบว่าอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าความถูกต้องสูงที่สุดอยู่ที่ 99.33%

เจษฎาพร ปาคำวัง, กาญจน คุ้มทรัพย์ และ วรชัย ศรีเมือง การเปรียบเทียบแบบจำลองจำแนกประเภทเพื่อการพัฒนาการท่องเที่ยว อำเภอเขาค้อ จังหวัดเพชรบูรณ์ด้วยเทคนิคเหมืองข้อมูล โดยใช้อัลกอริทึม Bayes network classifier,Naive bayes,Logistic regression,Sequential minimal optimization,Decision tree และ Random forest จากการทดสอบประสิทธิภาพอัลกอริทึมทั้ง 7 อัลกอริทึม พบว่า Decision tree ให้ค่าความถูกต้องสูงที่สุดอยู่ที่ 94.57%

ณัฐวดี หงส์บุญมี และ ประภาสิริ ตรีพาณิชกุล การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลเพื่อวิเคราะห์ปัจจัยความเสี่ยงที่ส่งผลต่อการเกิดโรคไฮเปอร์ไทรอยด์ด้วยเทคนิคเหมืองข้อมูลโดยใช้อัลกอริทึมโครงข่ายประสาทเทียม, การเรียนรู้แบบเบย์และตันไม้ตัดสินใจ จากการทดสอบประสิทธิภาพอัลกอริทึมทั้ง 3 อัลกอริทึมม พบว่าอัลกอริทึมโครงข่ายประสาทเทียมให้ค่าความถูกต้องสูงที่สุดอยู่ที่ 82.97%

วิธีการวิจัย

โดยผู้วิจัยได้ดำเนินงานวิจัยวิเคราะห์ข้อมูลตามกระบวนการมาตรฐานในการทำเหมืองข้อมูล (Cross-Industry Standard Process for Data Mining หรือ CRISP-DM) โดยแบ่งออกเป็น 6ขั้นตอนดังนี้



ภาพที่ 2 ขั้นตอนการดำเนินการวิจัย

1. ขั้นตอนการทำความเข้าใจปัญหา (Business Understanding)

เห็ดเป็นสิ่งมีความหลากหลายและมีความสำคัญในเชิงทัศนคติและทางโภชนาการในสังคมของเรา แต่เราก็พบว่ามีปัญหาจำนวนมากที่เกี่ยวข้องกับเห็ด ตั้งแต่ความไม่แน่นอนในความปลอดภัยของการบริโภคจนถึงการระบุชนิดของเห็ดอย่างถูกต้องและการเข้าใจสมบัติทางโภชนาการของเห็ด ดังนั้น การใช้เทคนิคการทำเหมืองข้อมูลกับชุดข้อมูล Mushroom จากฐานข้อมูล UCI ปี 1987 เป็นเครื่องมือที่มีความสำคัญในการแก้ไขปัญหาที่เกี่ยวข้องกับเห็ดและการบริโภคเห็ดในสังคมของเราได้อย่างมีประสิทธิภาพและปลอดภัยขึ้น

2. ขั้นตอนการทำความเข้าใจข้อมูล (Data Understanding)

ชุดข้อมูลสำหรับเห็ดถูกเก็บรวบรวมมาจากที่เก็บข้อมูล UCI (UCI repository) ชุดข้อมูลนี้ประกอบด้วยตัวอย่างเห็ดจาก Agaricus และ Lepiota แล้วจัดหมวดหมู่ว่าเห็ดเหล่านี้เป็นรายการอาหารที่มีความปลอดภัยอย่างแน่นอน (definitely edible) หรือเป็นสารพิษอย่างแน่นอนหรือเป็นสารที่ไม่แน่ใจถึงความอาจเป็นอาหารได้และไม่แนะนำให้บริโภค (definitely poisonous or of unknown edibility and not recommended) ชุดข้อมูลนี้ประกอบด้วยตัวอย่างของเห็ดทั้งหมด 8,124 ตัวอย่าง และมีคุณสมบัติทั้งหมด 22 คุณสมบัติ มี 2 ป้ายกำกับคลาสคือ "e" (definitely edible) และ "p" (definitely poisonous or of unknown edibility and not recommended)

3. ขั้นตอนการเตรียมข้อมูล (Data Preparation)

การจัดเตรียมและคัดเลือกแอตทริบิวต์ที่มีความสำคัญต่อการนำไปวิเคราะห์ข้อมูลเพื่อใช้ใสำหรับการแยกประเภทข้อมูลเห็ดด้วยเทคนิคการท้าเหมืองข้อมูล มีขั้นตอนการเตรียมข้อมูลดังต่อไปนี้

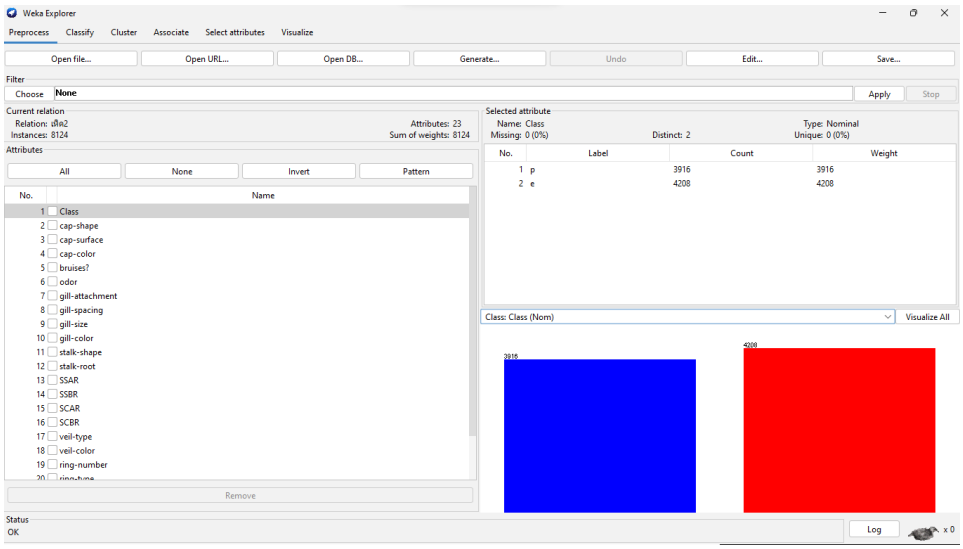
3.1 น้ำเข้าข้อมูลสำหรับการวิเคราะห์ที่เป็นข้อมูลดิบ (Raw Data) จากฐานข้อมูลUCI ให้อยู่ในรูปของ Excel หลังจากนำส่งออกข้อมูลมาเก็บในชนิดไฟล์ csv เพื่อเตรียมข้อมูลให้พร้อมสำหรับการประมวลผลในโปรแกรม Weka

3.2 คัดเลือกแอตทริบิวต์ที่ใช้ในการสร้างตัวแบบสำหรับการแยกประเภทข้อมูลเห็ดและในการทำความสะอาดข้อมูลไม่พบข้อมูลสูญหาย (Missing Data) ของทุกแอตทริบิวต์ในช่วงเวลาที่ศึกษา

3.3 กำหนดและตรวจสอบความถูกต้องของประเภทข้อมูล (Data Type) ที่จัดเก็บในแต่ละแอตทริบิวต์แสดงในตารางที่ 1

ตารางที่ 1 รายละเอียดชุดข้อมูล

ลำดับ	ชื่อ	ลักษณะข้อมูล	ตัวอย่างข้อมูล
1	Class	nominal	edible=e, poisonous=p
2	cap-shape	nominal	convex=x,flat=f
3	cap-surface	nominal	fibrous=f,grooves=g
4	cap-color	nominal	brown=n,buff=b
5	bruises?	nominal	bruises=t,no=f
6	odor	nominal	almond=a,anise=l
7	gill-attachment	nominal	free=f,notched=n
8	gill-spacing	nominal	close=c,crowded=w
9	gill-size	nominal	broad=b,narrow=n
10	gill-color	nominal	black=k,brown=n
11	stalk-shape	nominal	enlarging=e,tapering=t
12	stalk-root	nominal	bulbous=b,club=c
13	stalk-surface-above-ring	nominal	fibrous=f,scaly=y
14	stalk-surface-below-ring	nominal	silky=k,smooth=s
15	stalk-color-above-ring	nominal	brown=n,buff=b
16	stalk-color-below-ring	nominal	gray=g,orange=o
17	veil-type	nominal	partial=p,universal=u
18	veil-color	nominal	brown=n,orange=o
19	ring-number	nominal	one=o,two=t
20	ring-type	nominal	large=l, none=n
21	spore-print-color	nominal	black=k,brown=n
22	population	nominal	abundant=a,clustered=c
23	habitat	nominal	grasses=g,leaves=l



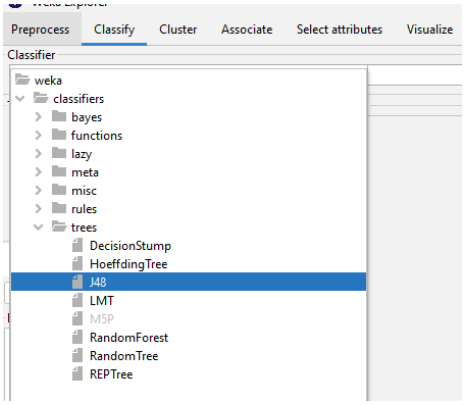
ภาพที่ 3 แสดงรายละเอียดเกี่ยวกับข้อมูลใน Weka

4. ขั้นตอนการสร้างตัวแบบ (Modeling)

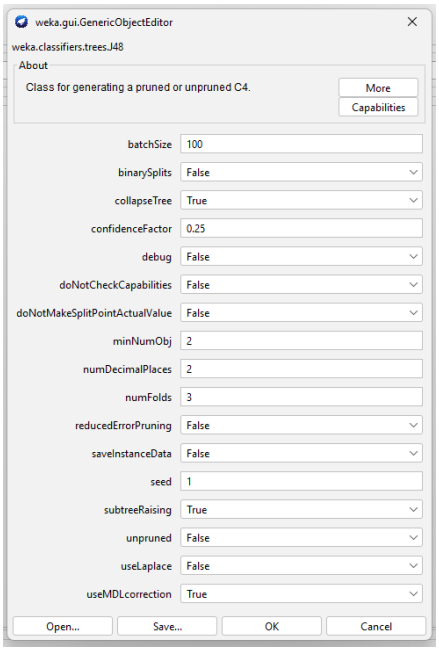
ในส่วนการสร้างโมเดลการจัดประเภท จะอาศัยข้อมูลในส่วนของชุดข้อมูลที่ ได้จัดทำไว้แล้วมาเป็นส่วนของข้อมูลสอน ซึ่งส่วนของการสร้างโมเดลจัดประเภทจะใช้อัลกอรึทึมวิธีต่างๆ ในการทดสอบกับชุดข้อมูลสอน โดยนำชุด ข้อมูลสอนที่สกัดได้จำนวน 8,124 ชุดข้อมูล ที่ผ่านกระบวนการทำความสะอาดข้อมูลรียบร้อยแล้วเข้าสู่ โปรแกรมWEKA เพื่อไปทดสอบการสร้างโมเดลการจำแนกประเภทคั่วยอัลกอริทึมประเภทต่าง ๆ โดยมีขั้นตอนการสร้างตัวจำแนกแต่ละอัลกอริทึมดังนี้

ต้นไม้ตัดสินใจ

จากข้อมูลสอนจำนวน 8,124 ชุด ที่ได้เตรียมไว้นำมาสร้างตัวจำแนกประเภทโดยใช้อัลกอริทึมต้นไม้ตัดสินใจ แบบ J48ด้วยโปรแกรมWeka แสดงดังภาพที่ 4 สำหรับการกำหนดค่าการสร้างตัวจำแนกประเภทดังภาพที่ 5



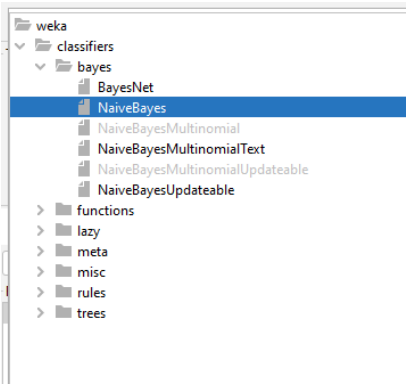
ภาพที่ 4 การสร้างตัวจำแนกประเภทด้วยอัลกอริทึมต้นไม้ตัดสินใจ (J48)



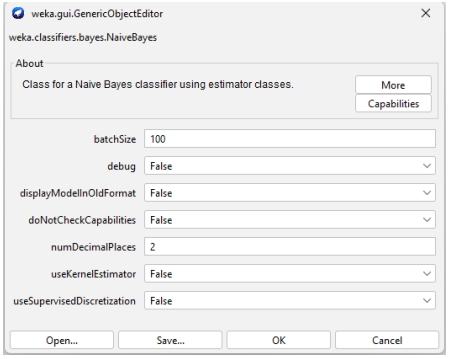
ภาพที่ 5 การกำหนดค่าในการสร้างอัลกอริทึมต้นไม้ตัดสินใจ (J48)

การเรียนรู้แบบเบย์

จากข้อมูลสอนจำนวน 8,124 ชุด ที่ได้เตรียมไว้นำมาสร้างตัวจำแนกประเภทโดยใช้อัลกอริทึม Naive Bayes ด้วยโปรแกรม Weka ดังภาพที่ 6 และกำหนดค่าการสร้างตัวจำแนกประเภท ดังภาพที่ 7



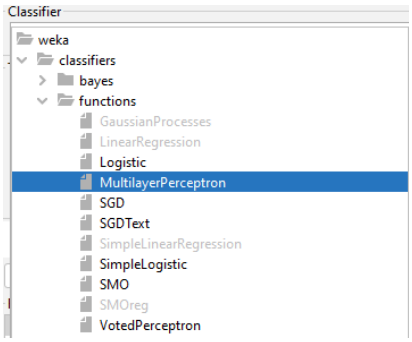
ภาพที่ 6 สร้างตัวจำแนกประเภทด้วย Naive Bayes



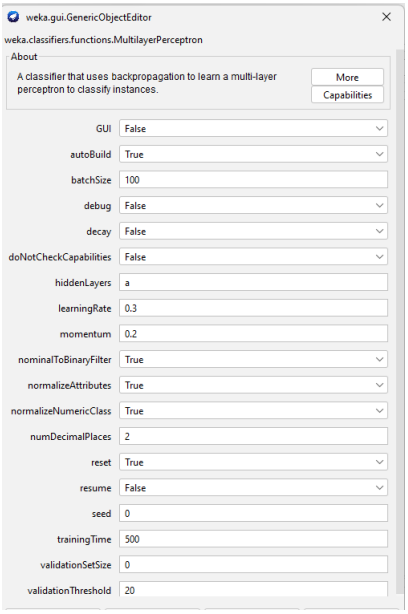
ภาพที่ 7 การกำหนดค่าในการสร้างอัลกอริทึม Naive Bayes

โครงข่ายประสาทเทียม

จากข้อมูลสอนจำนวน 8,124 ชุด ที่ได้เตรียมไว้นำมาสร้างตัวจำแนกประเภทโดยใช้อัลกอริทึม Artificial Neural network ด้วยโปรแกรม Weka ดังภาพที่ 8 และกำหนดค่าการสร้างตัวจำแนกประเภท ดังภาพที่ 9



ภาพที่ 8 สร้างตัวจำแนกประเภทด้วย โครงข่ายประสาทเทียม (Multilayer Perceptron)



ภาพที่ 9 การกำหนดค่าในการสร้างอัลกอริทึม โครงข่ายประสาทเทียม (Multilayer Perceptron)

5. ขั้นตอนการประเมินผล (Evaluation)

วิธีการมาตรฐานที่ใช้ในการประเมิน (สุวิทย์ กิระวิทยา, 2565) หรือวัดประสิทธิภาพของแบบจำลอง (model) นั้นมีหลากหลายวิธี วิธีการหนึ่งที่เป็นพื้นฐาน คือการแบ่งหมวดหมู่ของผลลัพธ์ โดยเราจะต้องมีข้อมูลผลลัพธ์จริงและผลลัพธ์ที่ได้จากแบบจำลองแล้วนำมาเทียบกันดู โดยในเรื่องการประเมินนี้ มีคำสำคัญที่จะต้องทราบความหมายหลัก ๆ คือ

1.ค่าความแม่นยำ (Precision): ค่านี้บ่งบอกถึงความแม่นยำของโมเดลในการทำนาย คือค่าที่วัดสัดส่วนของการทำนายที่ถูกต้องเมื่อเทียบกับที่ทำนายทั้งหมด สูตรคำนวณคือ $\frac{\mathrm{TP}}{\mathrm{TP+FP}}$

2.ค่าความระลึก (Recall): ค่านี้บ่งบอกถึงความสามารถในการตรวจจับทุกสิ่งที่เป็นข้อมูลที่จริง คือจำนวนที่ทำนายถูกต้องเมื่อเทียบกับจำนวนทั้งหมดของข้อมูลที่เป็นข้อมูลจริง สูตรคำนวณคือ $\frac{TP}{TP+FN}$

3.ค่าความถ่วงดุล (F-measure): ค่านี้เป็นการรวมค่าความแม่นยำและค่าความระลึกเข้าด้วยกันเพื่อให้ได้ค่าความถูกต้องที่ครอบคลุมทั้งการทำนายแม่นยำและการตรวจจับอย่างเท่าเทียม สูตรคำนวณคือ $\frac{(\text{Precision x Recall})}{(\text{Precision x Recall})}$

4.ค่าความถูกต้อง (Accuracy): ค่านี้บ่งบอกถึงความถูกต้องของโมเดลโดยรวม คือจำนวนข้อมูลที่ทำนายถูกต้องทุกคลาส เมื่อเทียบกับจำนวนข้อมูลทั้งหมด สูตรคำนวณคือ $\frac{\text{TP} + \text{TN}}{\text{TP+FP+FN+TN}}$

(ณัฐวดี และประภาสิริ, 2562)

TP = ข้อมูลเป็นจริงและผลการทำนายบอกว่าจริง

TN = ข้อมูลเป็นข้อมูลไม่จริงและผลการทำนายบอกว่าไม่จริง

FP = ข้อมูลเป็นจริงแต่ผลการทำนายบอกว่าไม่จริง

FN = ข้อมูลเป็นข้อมูลไม่จริงแต่ผลการทำนายบอกว่าจริง

```
=== Summary ===

Correctly Classified Instances        3250               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances             3250

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     p
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     e
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000
```

ภาพที่ 10 ตัวอย่างผลของอัลกอริทึมต้นไม้ตัดสินใจ (J48)

```
=== Summary ===

Correctly Classified Instances        3076               94.6462 %
Incorrectly Classified Instances       174                5.3538 %
Kappa statistic                          0.8927
Mean absolute error                      0.0522
Root mean squared error                  0.2019
Relative absolute error                 10.4438 %
Root relative squared error             40.3704 %
Total Number of Instances             3250

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.900    0.009    0.990      0.900    0.943      0.896    0.997     0.997     p
                 0.991    0.100    0.912      0.991    0.950      0.896    0.997     0.997     e
Weighted Avg.    0.946    0.055    0.950      0.946    0.946      0.896    0.997     0.997
```

ภาพที่ 11 ตัวอย่างผลของอัลกอริทึม Naive Bayes

```
Correctly Classified Instances        3250               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0002
Root mean squared error                  0.0005
Relative absolute error                  0.0345 %
Root relative squared error              0.1055 %
Total Number of Instances             3250

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     p
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     e
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

    a    b   <-- classified as
 1591    0 |    a = p
    0 1659 |    b = e
```

ภาพที่ 12 ตัวอย่างผลอัลกอริทึม โครงข่ายประสาทเทียม (Multilayer Perceptron)

6. ขั้นตอนการนำไปใช้งาน(Deployment)

การทำเหมืองข้อมูลเพื่อจำแนกประเภทข้อมูลเห็ดมีประโยชน์ในหลายด้านและสามารถนำไปใช้ในหลายหน่อยงานต่างๆ เพื่อเพิ่มประสิทธิภาพและความคุ้มค่าในการใช้ข้อมูลเห็ดของคุณในการตัดสินใจและการวิเคราะห์ต่าง ๆ ที่เกี่ยวข้องกับเห็ดและสิ่งที่เกี่ยวข้องกับมันในองค์กรหรืองานทางการแพทย์ ช่วยในการค้นหาข้อมูลเกี่ยวกับสมบัติของเห็ดเพื่อการวินิจฉัยและรักษา

ผลการวิจัยและวิจารณ์ผลการวิจัย

ผลการทดลองของงานวิจัยประกอบด้วย 4 ส่วน ได้แก่ 1) ผลการสร้างแบบจำลองด้วยต้นไม้ตัดสินใจ 2) ผลการสร้างแบบจำลองด้วยการเรียนรู้แบบเบย์ 3) ผลการสร้างแบบจำลองด้วยโครงข่ายประสาทเทียมและ 4) ผลการเปรียบเทียบประสิทธิภาพแบบจำลอง ดังนี้

1. ผลการสร้างแบบจำลองด้วยต้นไม้ตัดสินใจ

ตารางที่ 2 ค่าประสิทธิภาพจากเทคนิคต้นไม้ตัดสินใจ

RATIO TRAIN: TEST	ACCURACY (%)	PRECISION (%)	RECALL (%)	F-MEASURE (%)
60:40	100	100	100	100
70:30	100	100	100	100
80:20	100	100	100	100

2. ผลการสร้างแบบจำลองด้วยการเรียนรู้แบบเบย์

ตารางที่ 3 ค่าประสิทธิภาพจากเทคนิคการเรียนรู้แบบเบย์

RATIO TRAIN: TEST	ACCURACY (%)	PRECISION (%)	RECALL (%)	F-MEASURE (%)
60:40	94.64	99.20	91.20	95
70:30	95.40	92.40	99.30	95.70
80:20	95.38	95.60	95.40	95.40

3. ผลการสร้างแบบจำลองด้วยโครงข่ายประสาทเทียม

ตารางที่ 4 ค่าประสิทธิภาพจากเทคนิคโครงข่ายประสาทเทียม

RATIO TRAIN: TEST	ACCURACY (%)	PRECISION (%)	RECALL (%)	F-MEASURE (%)
60:40	100	100	100	100
70:30	100	100	100	100
80:20	100	100	100	100

4. ผลการเปรียบเทียบประสิทธิภาพแบบจำลอง

จากการเปรียบเทียบประสิทธิภาพการทำงานของทั้งสามเทคนิคได้ผลการทดลองดังตารางที่ 5 พบว่าแบบจำลองจากเทคนิคต้นไม้ตัดสินใจ มีค่าความถูกต้อง 100% ค่าความแม่นยำ 100% ค่าความระลึก 100% และค่าความถ่วงดุล 100% แบบจำลองจากเทคนิคการเรียนรู้แบบเบย์ มีค่าความถูกต้อง 95.40% ค่าความแม่นยำ 95.4% ค่าความระลึก 95.3%และค่าความถ่วงดุล 95.7% และแบบจำลองแบบโครงข่ายประสาทเทียม มีค่าความถูกต้อง 100% ค่าความแม่นยำ 100% ค่าความระลึก 100% และค่าความถ่วงดุล 100%

ตารางที่ 5 เปรียบเทียบประสิทธิภาพแบบจำลอง

Ratio Train: Test	*Accuracy* (%)	*Precision* (%)	*Recall* (%)	*F-measure* (%)
Decision Tree	100	100	100	100
Naïve Bayes	95.40	92.4	99.3	95.70
Neural Network	100	100	100	100

สรุปผลการวิจัย

งานวิจัยนี้เป็นการการแยกประเภทข้อมูลเห็ดด้วยเทคนิคการทำเหมืองข้อมูล ซึ่งมีวัตถุประสงค์เพื่อนำเทคนิคเหมืองข้อมูลแบบจำแนกประเภทข้อมูลเห็ดพร้อมทั้งเปรียบประสิทธิภาพ เพื่อให้ได้อัลกอริทึมที่มีประสิทธิภาพที่เหมาะสม โดยใช้ข้อมูล จากฐานข้อมูล UCI Mushroomปี1987 ซึ่งมีจำนวนข้อมูล 8,124 ชุดข้อมูล และปัจจัยสำหรับการวิเคราะห์ 22 ปัจจัย ซึ่งทำการนำข้อมูลของข้อมูลเห็ดที่จัดเก็บมาสร้างโมเดลในการทำนายโดยใช้ เทคนิค ต้นไม้ตัดสินใจ (Decision Tree) การเรียนรู้แบบเบย์(Naive Bayes) และโครงข่ายประสาทเทียม (Artificial Neural network) ผ่านกระบวนการขั้นตอนการดำเนินงานวิจัยโดยการประยุกต์ใช้โดยกระบวนการมาตรฐานในการทำเหมืองข้อมูล (Cross-Industry Standard Process for Data Mining หรือ CRISP-DM)มาเป็นแนวทางในการดำเนินงานวิจัยและการแบ่งข้อมูล Training ที่ใช้ในการทดลองมีการแบ่ง 60:40, 70:30, 80:20 ผลการวิจัยพบว่า ตัวแบบจำลองการจำแนกประเภทข้อมูลด้วย ต้นไม้ตัดสินใจ (Decision Tree) กับโครงข่ายประสาทเทียม (Artificial Neural network) มีประสิทธิภาพเท่ากัน โดยมีค่าความถูกต้อง 100% ค่าความแม่นยำ 100% ค่าความระลึก 100% ค่าความถ่วงดุล 100% และ การเรียนรู้แบบเบย์(Naive Bayes) มีค่าต่ำสุดของทั้ง3อัลกอริทึม โดยมีค่าความถูกต้อง 92.40% ค่าความแม่นยำ 95.40% ค่าความระลึก 99.30% ค่าความถ่วงดุล 95.70% ดังนั้นจึงสรุปได้ว่า

เทคนิค ต้นไม้ตัดสินใจ (Decision Tree) และ โครงข่ายประสาทเทียม (Artificial Neural network) มีความเหมาะสมในการนำมาสร้างแบบจำลองการจำแนกประเภทข้อมูลของเห็ด

ข้อเสนอแนะ

ในการทำวิจัยครั้งต่อไป ควรมีวิธีการแบ่งชุดข้อมูลที่หลากหลายรูปแบบสำหรับการศึกษา ฝึกฝนและการทดสอบ อีกทั้งควรมีการนำอัลกอริทึมหลายแบบมาทำการทดลอง เช่นซัพพอร์ตเวกเตอร์แมชชีน(Support Vector Machines) เพื่อหาประสิทธิภาพในการจำแนกหรือการทำนายที่เหมาะสมกว่า

กิตติกรรมประกาศ

ขอบคุณแหล่งข้อมูลเปิด Machine Learning Repository จาก Algerian Forest Fires Dataset (UCI)

อ้างอิง

ธนาวุฒิ ประกอบผล. (2552).**โครงข่ายประสาทเทียม (Artificial Neural network)**. วารสารมหาวิทยาลัยหัวเฉียวเฉลิมพระเกียรติ, 73-87.

วทัญญู ชูประจิตต์ [1], ผู้ช่วยศาสตราจารย์ดร.นพมาศ ปักเข็ม [2] และ ผู้ช่วยศาสตราจารย์ดร.สิรยา สิทธิสาร [3]. (2552). **การจำแนกประเภทและวิธีการคัดเลือกคุณลักษณะสำหรับการสนับสนุนการวิเคราะห์ความรู้สึกต่อสถานที่ท่องเที่ยวไทย**. วิทยานิพนธ์มหาวิทยาลัยทักษิณ, 44-49.

เจษฎาพร ปาคำวัง, กาญจน คุ้มทรัพย์ และ วรชัย ศรีเมือง. (2563). **การเปรียบเทียบแบบจำลองจำแนกประเภทเพื่อการพัฒนาการท่องเที่ยว อำเภอเขาค้อ จังหวัดเพชรบูรณ์**. Rajabhat J. Sci. Humanit. Soc. Sci, 21(1), 213-223.

สุวิทย์ กิระวิทยา. (2565). **การประเมินแบบจำลองจากการแบ่งหมวดหมู่ (Evaluation Model by Classification)**. http://suwitkiravittaya.eng.chula.ac.th/suwitnote/Model_Evaluation.htm

ณัฐวดี หงส์บุญมี และ ประภาสิริ ตรีพาณิชกุล. (2563). **การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลเพื่อวิเคราะห์ปัจจัยความเสี่ยงที่ส่งผลต่อการเกิดโรคไฮเปอร์ไทรอยด์ด้วยเทคนิคเหมืองข้อมูล**. Journal of Information Science and Technology, 9(1), 41-51.

จิรโรจน์ ตอสะสุกุล*และสุพิชชา ขัดธิพงษ์[2]. (2565). **การเปรียบเทียบประสิทธิภาพของตัวแบบการพยากรณ์ ปริมาณน้ำในเขื่อนด้วยเทคนิคการท้าเหมืองข้อมูล.** วารสารการบัญชีและการจัดการ มหาวิทยาลัยมหาสารคาม

Natdanai James. (2019). **Supervised Learning การเรียนรู้ของเครื่องแบบมีผู้สอน**. https://www.glurgeek.com/education/ie311supervisedlearning/

Rattanatat. (2019). **การทำเหมืองข้อมูล (Data Mining)**. กรมส่งเสริมอุตสาหกรรรม กองโลจิสติกส์. https://dol.dip.go.th/th/category/2019-02-08-08-57-30/2019-03-15-08-49-57

รุจิรา ธรรมสมบัติ. (2554).**ระบบสนับสนุนการตัดสินใจในการเลือกใช้แพคเกจอินเทอร์เน็ตมือถือ โดยใช้ต้นไม้ตัดสินใจ**. มหาวิทยาลัยราชพฤกษ์.

สถาบันนวัตกรรมและธรรมาภิบาลข้อมูล. (2565) **Naive Bayes Classification**. https://digi.data.go.th/blog/what-is-classification-model/