

การเปรียบเทียบประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่อง

ในการจำแนกแหล่งกำเนิดทางภูมิศาสตร์ของมรกต

Comparison of Machine Learning Model Performance in Classifying the Geographical Origin of Emeralds

ชาวลิต จันคันธา^{1*} และ รัฐพล อำพล²

Chawalit Chankhantha^{1*} and Rattaphon Amphon²

¹กองพัฒนาอุตสาหกรรมสร้างสรรค์ กรมส่งเสริมอุตสาหกรรม

²ภาควิชาธรณีวิทยา คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

¹*Division of Creative Industry Development, Department of Industrial Promotion*

²*Department of Geological Sciences, Faculty of Science, Chiang Mai University*

E-Mail: charles.valydh@gmail.com

Received: November 14, 2025; Revised: June 26, 2026; Accepted: June 26, 2026; Published: June 30, 2026

ABSTRACT – Currently, the classification of emerald origins is of significant commercial importance, as it directly impacts the value and credibility of the gemstones. However, traditional identification methods often encounter challenges due to the similarity in chemical compositions across different origins, leading to analytical inaccuracies. Therefore, this research aims to develop machine learning models to classify the origin of emeralds and to compare the performance of each model. The research methodology followed the 6 phases of the Cross-Industry Standard Process for Data Mining (CRISP-DM). The dataset utilized in this study consists of trace element data of emeralds from 4 origins, totaling 1088 samples: Colombia, Zambia (Kafubu), Zambia (Musakashi), and Afghanistan. The dataset was partitioned into training and testing sets at an 80:20 ratio. Subsequently, feature selection was performed utilizing One-Way ANOVA and Pearson correlation coefficient techniques, resulting in 10 optimal features. To construct the models, 6 machine learning algorithms were selected: Random Forest, Support Vector Machine, Naïve Bayes, Logistic Regression, Linear Discriminant Analysis, and k-Nearest Neighbors. The results revealed that the Random Forest model yielded the highest performance in classifying emerald origins, achieving an accuracy, precision, recall, and F1 Score of 99.54%, 99.55%, 99.54%, and 99.54%, respectively. These findings demonstrate that machine learning techniques can be applied as highly effective tools for accurately classifying gemstone origins, significantly enhancing confidence in gemstone analysis.

KEY WORDS -- Emerald, Machine Learning, Geographical Origin, Model, Classification

บทคัดย่อ -- ปัจจุบันการจำแนกแหล่งกำเนิดของมรกตมีความสำคัญอย่างยิ่งในการค้าอัญมณี เนื่องจากส่งผลโดยตรงต่อมูลค่าและความน่าเชื่อถือของอัญมณี อย่างไรก็ตามพบว่าการตรวจสอบแบบเดิมมักประสบปัญหาความคลาดเคลื่อนขององค์ประกอบทางเคมีของมรกตจากแหล่งต่าง ๆ ทำให้เกิดความคลาดเคลื่อนในการวิเคราะห์ ดังนั้นงานวิจัยนี้จึงมีวัตถุประสงค์เพื่อพัฒนาแบบจำลองการเรียนรู้ของเครื่องในการจำแนกแหล่งกำเนิดของมรกต และเปรียบเทียบประสิทธิภาพของแบบจำลองแต่ละประเภท การดำเนินงานวิจัยใช้กระบวนการมาตรฐานสำหรับการทำเหมืองข้อมูล (CRISP-DM)

จำนวน 6 ขั้นตอน ชุดข้อมูลที่ใช้ในการศึกษาเป็นข้อมูลธาตุรองรอยของมรกตจาก 4 แหล่ง จำนวน 1088 ตัวอย่าง ได้แก่ โคลอมเบีย แซมเบีย (คาฟู) แซมเบีย (มุซาคาซี) และอัฟกานิสถาน ชุดข้อมูลถูกแบ่งออกเป็นข้อมูลฝึกสอนและข้อมูลทดสอบในอัตราส่วน 80:20 จากนั้นทำการคัดเลือกคุณลักษณะด้วยเทคนิค One-Way ANOVA และค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน ได้คุณลักษณะที่เหมาะสมจำนวน 10 ตัวแปร สำหรับการสร้างแบบจำลอง การพัฒนาแบบจำลองการเรียนรู้ของเครื่องจำนวน 6 ประเภท ได้แก่ ป่าสุ่ม ซัพพอร์ตเวกเตอร์แมชชีน นาอ์ฟเบย์ การถดถอยโลจิสติกส์ การวิเคราะห์จำแนกประเภทเชิงเส้น และเพื่อนบ้านใกล้ที่สุด-เค ผลการวิจัยพบว่า แบบจำลองป่าสุ่มให้ประสิทธิภาพสูงสุดในการจำแนกแหล่งกำเนิดของมรกต โดยมีค่าความถูกต้อง ความแม่นยำ ความความระลึก และค่า F1 Score เท่ากับร้อยละ 99.54 99.55 99.54 และ 99.54 ตามลำดับ จากผลการวิจัยแสดงให้เห็นว่าเทคนิคการเรียนรู้ของเครื่องสามารถนำมาใช้เป็นเครื่องมือที่มีประสิทธิภาพในการจำแนกแหล่งกำเนิดของอัญมณีได้อย่างแม่นยำ และสามารถช่วยเพิ่มความเชื่อมั่นในการวิเคราะห์อัญมณีได้อย่างมีนัยสำคัญ

คำสำคัญ -- มรกต, การเรียนรู้ของเครื่อง, แหล่งกำเนิดทางภูมิศาสตร์, แบบจำลอง, การจำแนก

1. บทนำ

อุตสาหกรรมอัญมณีและเครื่องประดับมีบทบาทสำคัญในการขับเคลื่อนเศรษฐกิจของประเทศ เนื่องจากเป็นกลุ่มสินค้าส่งออกสำคัญที่อยู่ในอันดับต้น ๆ ติดต่อกันมาหลายปี ปัจจุบันการระบุแหล่งกำเนิดของอัญมณี (Origin Determination of Gemstones) มีความสำคัญต่อการค้าขายอัญมณีเป็นอย่างมาก เนื่องจากแหล่งที่มาส่งผลต่อมูลค่าของอัญมณี โดยเฉพาะกลุ่มอัญมณีที่มีความต้องการสูง เช่น ทับทิมจากเหมืองโมกอก ประเทศเมียนมาร์ มักมีราคาสูงกว่าทับทิมจากแหล่งอื่นหลายเท่า เมื่อเทียบน้ำหนักและคุณภาพใกล้เคียงกัน จึงเป็นความท้าทายของนักวิเคราะห์อัญมณีที่ต้องอาศัยประสบการณ์ความเชี่ยวชาญสูง ในช่วงปี ค.ศ. 1950 เริ่มมีการระบุแหล่งกำเนิดในใบรับรองอัญมณีโดยห้องปฏิบัติการระดับโลก ได้แก่ Gübelin Gem Lab และ Schweizerische Stiftung für Edelstein-Forschung (SSEF) ประเทศสวิตเซอร์แลนด์ ซึ่งในขณะนั้นการค้นพบแหล่งอัญมณียังไม่มากเท่าปัจจุบัน เช่น แหล่งมรกตส่วนใหญ่จะมาจากประเทศโคลอมเบีย และรัสเซีย เป็นต้น [1] การระบุแหล่งกำเนิดของอัญมณียังเป็นกลไกที่ช่วยสร้างความโปร่งใสและตรวจสอบย้อนกลับได้ (Traceability) ตลอดห่วงโซ่อุปทาน สอดคล้องกับข้อตกลงทางการค้าระหว่างประเทศในด้านความรับผิดชอบต่อสังคมและสิ่งแวดล้อม โดยมีองค์กร Responsible Jewellery Council (RJC) ทำหน้าที่กำกับดูแล

เพื่อให้มั่นใจว่าอัญมณีแหล่งที่มาที่ถูกต้องตามกฎหมาย มาจากธุรกิจที่ไม่เกี่ยวข้องกับการละเมิดสิทธิมนุษยชน และไม่ส่งผลกระทบต่อสิ่งแวดล้อม นอกจากนี้ยังสร้างความมั่นใจให้กับผู้บริโภคอีกด้วย [2] ซึ่งทั้งหมดนี้เป็นรากฐานสำคัญในการพัฒนาอุตสาหกรรมอัญมณีและเครื่องประดับให้เติบโตอย่างยั่งยืน

ในการจำแนกแหล่งกำเนิดของมรกตนั้นต้องอาศัยทั้งประสบการณ์และความเชี่ยวชาญของนักอัญมณี โดยทั่วไปจะพิจารณาจากข้อมูลเอกลักษณ์ทางเคมี ข้อมูลทางสเปกโทรสโกปี เช่น UV-Vis-NIR Spectroscopy และลักษณะเฉพาะของมลทินภายใน (Inclusions) เพื่อให้ผลที่แม่นยำมากขึ้น [3] แต่เนื่องจากมรกตเกิดในสภาพแวดล้อมทางธรณีวิทยาคล้ายคลึงกัน บ่อยครั้งที่เจอพบข้อมูลมีความใกล้เคียงกัน ส่งผลต่อความแม่นยำในการตัดสินใจ จากปัญหาดังกล่าว ผู้วิจัยเล็งเห็นถึงความสำคัญในการพัฒนาวิธีการจำแนกแหล่งกำเนิดของมรกตให้มีประสิทธิภาพ โดยนำเทคนิคการเรียนรู้แบบมีผู้สอน (Supervised Learning) ของ ปัญญาประดิษฐ์ (Artificial Intelligence) มาประยุกต์ใช้กับข้อมูลผลวิเคราะห์องค์ประกอบทางเคมีของมรกตจากแหล่งที่มีศักยภาพในเชิงพาณิชย์ ทำการฝึกฝนและทดสอบข้อมูลผ่านแบบจำลองการเรียนรู้ของเครื่องจำนวน 6 ประเภท ได้แก่ ป่าสุ่ม (Random Forests: RF) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machines: SVM) นาอ์ฟเบย์ (Naïve Bayes: NB) การถดถอยโลจิสติกส์ (Logistic

Regression: LR) การวิเคราะห์การจำแนกเชิงเส้น (Linear Discriminant Analysis: LDA) และเพื่อนบ้านใกล้ที่สุด-เค (k-Nearest Neighbors: k-NN) การเลือกใช้แบบจำลองที่มีประสิทธิภาพที่สุด จะเป็นเครื่องมือสนับสนุนนักอัญมณีให้สามารถจำแนกแหล่งกำเนิดของมรกตได้อย่างถูกต้อง แม่นยำ ลดความผิดพลาดในการสำแดงแหล่งที่มา และยกระดับความน่าเชื่อถือในอุตสาหกรรมอัญมณีและเครื่องประดับ ทั้งนี้ ผลการวิจัยยังสามารถนำไปพัฒนาต่อยอดเทคนิคการเรียนรู้ของเครื่องเพื่อจำแนกแหล่งที่มาของอัญมณีชนิดอื่น ๆ ที่มีมูลค่าสูงในอนาคตได้อีกด้วย

2. เอกสารและงานวิจัยที่เกี่ยวข้อง

มรกต (Emerald) เป็นอัญมณีสีเขียว อยู่ในสปีชีส์เบริล (Beryl species) มีสูตรเคมี $\text{Be}_3\text{Al}_2\text{Si}_6\text{O}_{18}$ แสดงผลึกในระบบเฮกซะโกนอล (Hexagonal System) มีความแข็ง 7-7.5 ตามมาตราวัดของโมส์ โครงสร้างผลึกประกอบด้วยโครงข่ายของ SiO_4 , BeO_4 และ AlO_6 มีลักษณะคล้ายแหวนเรียงซ้อนกันเป็นชั้น ๆ จนเกิดเป็นท่อ (Chanel Sites) ขนานกันตามแนวแกนผลึก ท่อดังกล่าวนี้มีเส้นผ่านศูนย์กลางขนาด 2.80-5.10 อังสตรอม สามารถกักเก็บอัลคาไลไอออนขนาดเล็กจนถึงโมเลกุลใหญ่ เช่น Na^+ , Li^+ , K^+ , Rb^+ , Cs^+ , H_2O , CO_2 และก๊าซเฉื่อยบางชนิด [4] สีเขียวของมรกตเกิดจากธาตุโครเมียม (Cr) และวานาเดียม (V) เข้าไปแทนที่อะลูมิเนียม (Al) ในโครงสร้าง ส่วนธาตุเหล็ก (Fe) หากมีปริมาณสูงจะทำให้มรกตมีสีเข้ม เช่น มรกตจากประเทศแซมเบียที่มีสีเขียวเข้ม การรวมกันของธาตุเหล่านี้ส่งผลให้มรกตในธรรมชาติมีเจดสีตั้งแต่เขียวอ่อน ถึงเขียวเข้ม [5] โดยธรรมชาติมรกตมักมีรอยแตกร้าวภายในเนื้อพลอยอยู่แล้ว ดังนั้นจึงมักมีการปรับปรุงคุณภาพด้วยการอุดรอยแตกร้าวโดยใช้วัสดุที่มีค่าดัชนีหักเหของแสงใกล้เคียงกับมรกต (1.577-1.583) เช่น น้ำมันซิดาร์วูด เรซิน และพอลิเมอร์ โดยทั่วไปการอุดรอยแตกร้าวด้วยน้ำมันซิดาร์วูดได้รับการยอมรับมากกว่าการอุดด้วยเรซิน มรกตที่มีการค้าขายในปัจจุบันส่วนใหญ่มาจากประเทศโคลอมเบีย บราซิล แซมเบีย รัสเซีย มาดากัสการ์ และเอธิโอเปีย โดยเฉพาะมรกตจากเหมืองมุโซ ประเทศโคลอมเบีย ที่มีชื่อเรียกทางการค้าว่า Muzo Green เป็นมรกตคุณภาพสูง มีเนื้อค่อนข้างใส มีสีเขียวสด เป็นที่ต้องการและมี

ชื่อเสียงมาอย่างยาวนาน โดยในปี 2024 มรกตจากโคลอมเบียที่มีชื่อว่า The Aga Khan Emerald น้ำหนัก 37 กระรัต ถูกประมูลไปในราคาสูงถึง 305 ล้านบาท

ในด้านธรณีวิทยา มรกตเป็นอัญมณีที่เกิดขึ้นได้ค่อนข้างยากในธรรมชาติ เนื่องจากธาตุเบริลเลียม (Be) ที่เป็นองค์ประกอบหลักของมรกต พบได้น้อยมากในเปลือกโลก โดยมีความเข้มข้นเฉลี่ยเพียง 2.1 ส่วนในล้านส่วนโดยน้ำหนัก (ppmw) มรกตส่วนใหญ่เกิดในหินเพกมาไทต์ซึ่งมีธาตุเบริลเลียม อะลูมิเนียม และซิลิกา เข้าไปสัมผัสและทำปฏิกิริยาเคมีกับหินภูเขาไฟชนิดเมฟิก-อัลตราเมฟิก ซึ่งมีธาตุโครเมียม และวานาเดียม ในสภาวะที่อุณหภูมิและความดันเหมาะสมจนเกิดการตกผลึกของมรกตขึ้นมา อย่างไรก็ตามมรกตจากประเทศโคลอมเบียมีกระบวนการเกิดที่แตกต่างจากแหล่งอื่น โดยเกิดจากกระบวนการหมุนเวียนของน้ำแร่ร้อน (Hydrothermal Fluid) ภายในหินดินดานสีดำ (Black Shales) ในช่วงอุณหภูมิ 330–450 องศาเซลเซียส และความดันประมาณ 2–3.5 กิโลบาร์ [6]

การเรียนรู้ของเครื่อง (Machine Learning) เป็นสาขาย่อยของปัญญาประดิษฐ์ (Artificial Intelligence) ที่มุ่งพัฒนาเทคนิคและวิธีการให้คอมพิวเตอร์สามารถเรียนรู้จากการป้อนข้อมูลให้ระบบประมวลผลและค้นหาความสัมพันธ์ระหว่างข้อมูลเข้ากับข้อมูลขาออกเพื่อสร้างขั้นตอนการประมวลผลหรือแบบจำลองสำหรับใช้กับข้อมูลใหม่ในอนาคต การเรียนรู้ของเครื่องแบ่งออกเป็น 3 ประเภทหลัก ได้แก่ การเรียนรู้แบบมีผู้สอน (Supervised Learning) การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) และการเรียนรู้แบบเสริมแรง (Reinforcement Learning) [7]

การประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่องเพื่อจำแนกแหล่งกำเนิดทางภูมิศาสตร์ของอัญมณีได้มีการศึกษาและรายงานว่ามีความแม่นยำในระดับสูง เช่น การศึกษาของ Blodgett and Shen [8] รายงานการประยุกต์ใช้เทคนิคการวิเคราะห์การจำแนกกลุ่มเชิงเส้น (Linear Discriminant Analysis: LDA) ร่วมกับข้อมูลธาตุร่องรอย (Trace element) ที่ได้จากการวิเคราะห์ด้วยเทคนิค LA-ICP-MS เพื่อจำแนกแหล่งกำเนิดของอัญมณี ได้แก่ ทับทิม ไพลิน และพาราอับัว รวมถึงประยุกต์ใช้ในการตรวจแยกเพชรสีที่ผ่านการปรับปรุง

คุณภาพด้วยกระบวนการ High Pressure High Temperature (HPHT) ออกจากเพชรสังเคราะห์ ซึ่งผลการวิเคราะห์แสดงให้เห็นถึงความแม่นยำในการจำแนกได้เป็นอย่างดี

Luo et al. [9] จำแนกแหล่งกำเนิดของหยกเนไฟรต์จาก 8 แหล่งสำคัญในภูมิภาคเอเชียตะวันออกเฉียงใต้ โดยใช้ข้อมูลของธาตุจำนวน 8 ชนิด ได้แก่ Ti, V, Cr, Mn, Fe, Co, Ni และ Cu ร่วมกับเทคนิคการวิเคราะห์จำแนกกลุ่มเชิงเส้นแบบทวิภาควนซ้ำ (Iterative Binary Linear Discriminant Analysis) ผลการศึกษาพบว่าแบบจำลองดังกล่าวให้ความถูกต้องเฉลี่ยสูงถึงร้อยละ 94.40

Homkrajae et al. [10] ศึกษาเมทริกซ์จำนวน 225 ตัวอย่าง ซึ่งประกอบด้วยเมทริกซ์แร่จากประเทศจีน เมทริกซ์แร่ และเมทริกซ์จากสหรัฐอเมริกา โดยใช้เทคนิค LA-ICP-MS เพื่อวิเคราะห์ธาตุร่องรอย และทำการจำแนกด้วยเทคนิค LDA พบว่าแบบจำลองสามารถจำแนกประเภทของเมทริกซ์ได้ถูกต้องที่ร้อยละ 85 ซึ่งแสดงให้เห็นถึงศักยภาพของการใช้ข้อมูลธาตุร่องรอยร่วมกับการวิเคราะห์เชิงสถิติขั้นสูง

Wang and Krzemnicki [11] ประยุกต์ใช้เทคนิค LA-ICP-TOF-MS ในการวิเคราะห์องค์ประกอบธาตุของมรกตจำนวน 168 ตัวอย่าง และใช้เทคนิค t-Distributed Stochastic Neighbor Embedding (t-SNE) ซึ่งเป็นอัลกอริทึมการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) เพื่อจำแนกแหล่งกำเนิดของมรกต ผลการวิจัยชี้ว่า t-SNE ให้ประสิทธิภาพสูงกว่าเทคนิคการวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis: PCA) ในการจำแนกกลุ่มข้อมูล

Zhong et al. [12] ศึกษาการจำแนกแหล่งกำเนิดของเพริโดต (Peridot) จำนวน 262 ตัวอย่างจากสาธารณรัฐประชาชนจีนและเกาหลีเหนือ โดยใช้ข้อมูลทางเคมีจากการวิเคราะห์ด้วยเทคนิค LA-ICP-MS เพื่อพัฒนาแบบจำลองการเรียนรู้ของเครื่อง 6 ประเภท ได้แก่ LDA, Decision Tree, Random Forests, k-Nearest Neighbors (k-NN), Support Vector Machines และ Logistic Regression ในการศึกษาได้แบ่งข้อมูลออกเป็นชุดฝึกสอนและชุดทดสอบในอัตรา 70:30 ผลการศึกษาพบว่าแบบจำลอง SVM ให้ค่าความถูกต้องสูงสุด

นอกจากนี้ Seneewong-Na-Ayutthaya et al. [13] ประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่องเพื่อจำแนกแหล่งกำเนิด

ของคอร์นดัม โดยใช้ข้อมูลองค์ประกอบทางเคมีที่ได้จากการวิเคราะห์ด้วยเทคนิค LA-ICP-MS ผลการศึกษาพบว่าแบบจำลองป่าสุ่ม (Random Forests) สามารถจำแนกแหล่งกำเนิดของคอร์นดัมได้อย่างมีประสิทธิภาพ โดยให้ค่าความถูกต้อง ร้อยละ 82.50 ในกลุ่มทับทิม และร้อยละ 92.50 ในกลุ่มไพลิน อย่างไรก็ตามการใช้ข้อมูลเคมีจากเทคนิค EDXRF ให้ความแม่นยำน้อยกว่า

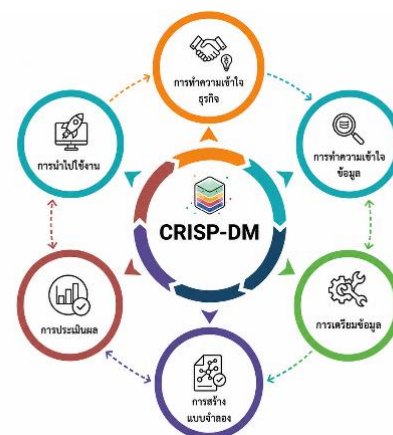
3. วิธีดำเนินการวิจัย

3.1 เครื่องมือการวิจัย

งานวิจัยนี้ใช้เครื่องมือ Google Collaboratory บนเว็บไซต์ <https://colab.research.google.com> ซึ่งเขียนด้วยภาษาไพธอน ซึ่งสามารถเข้าถึง GPU และรองรับการประมวลผลข้อมูล และการเรียนรู้ของเครื่องผ่านการทำงานจากไลบรารีต่าง ๆ ได้

3.2 ขั้นตอนการดำเนินการวิจัย

งานวิจัยนี้นำกระบวนการมาตรฐานสำหรับการทำเหมืองข้อมูล (Cross-Industry Standard Process for Data Mining: CRISP-DM) มาประยุกต์ใช้ ซึ่งประกอบ 6 ขั้นตอน [14] ดังแสดงในรูปที่ 1



รูปที่ 1. การดำเนินงานวิจัยตามกระบวนการของ CRISP-DM 6 ขั้นตอน

3.2.1 การทำความเข้าใจธุรกิจ

การระบุแหล่งกำเนิดของมรกตได้อย่างถูกต้องมีความสำคัญมากในทางการค้าเนื่องจากส่งผลโดยตรงต่อมูลค่าความโปร่งใส และความน่าเชื่อถือ ปัจจุบันอาศัยข้อมูลทางเคมี

และความเชี่ยวชาญของนักวิเคราะห์ข้อมูลเป็นผู้ชี้ขาด แต่มักประสบปัญหาบางแหล่งมีข้อมูลที่คล้ายคลึงกัน ทำให้ไม่สามารถระบุแหล่งได้ชัดเจน งานวิจัยนี้จึงมุ่งพัฒนาแบบจำลองการเรียนรู้ของเครื่องแบบมีผู้สอน จำนวน 6 ประเภท ได้แก่ การวิเคราะห์จำแนกประเภทเชิงเส้น (Linear Discriminant Analysis: LDA) การถดถอยโลจิสติกส์ (Logistic Regression) นาอิวเบย์ (Naïve Bayes) เพื่อนบ้านใกล้ที่สุด-เค (k-Nearest Neighbor, k-NN) ป่าสุ่ม (Random Forest) และซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine, SVM) เพื่อพัฒนาแบบจำลองที่มีประสิทธิภาพสูงสุดในการจำแนกแหล่งกำเนิดของมรดกได้อย่างถูกต้องและแม่นยำ

3.2.2 การทำความเข้าใจข้อมูล

รวบรวมชุดข้อมูลผลการวิเคราะห์ธาตุร่องรอย (Trace elements) ของมรดกที่ได้จากการวิเคราะห์ด้วยเทคนิค Laser Ablation Inductively Coupled Plasma Mass Spectrometry (LA-ICP-MS) [15] ประกอบด้วยตัวแปรตาม (Dependent variable) ซึ่งเป็นแหล่งมรดกจำนวน 4 แหล่ง ได้แก่ Afghanistan (116 ตัวอย่าง) Colombia (306 ตัวอย่าง) Zambia (Kafubu) (393 ตัวอย่าง) และ Zambia (Musakashi) (273 ตัวอย่าง) ส่วนตัวแปรอิสระ (Independent variables) เป็นข้อมูลเชิงปริมาณที่แสดงค่าความเข้มข้นของธาตุร่องรอยแต่ละชนิด โดยวัดค่าออกมาในหน่วยส่วนในล้านส่วนโดยน้ำหนัก (Parts per million by weight: ppmw)

3.2.3 การเตรียมข้อมูล

1) การทำความสะอาดและติดป้ายกำกับข้อมูล

ตรวจสอบข้อมูลที่สูญหาย และแทนค่าสูญหายดังกล่าวด้วยวิธีแทนค่าหลายค่า (Imputation) ปรับข้อมูลให้สมมาตรด้วยวิธี Min-max normalization เพื่อให้สามารถนำข้อมูลเข้าสู่กระบวนการวิเคราะห์ได้ ทำการลบข้อมูลในบางคอลัมน์ที่ไม่เกี่ยวข้อง จากนั้นทำการจัดการข้อมูลที่ไม่สมดุลด้วยวิธี SMOTE (Synthetic Minority Over-Sampling Technique) โดยอาศัยหลักการของเพื่อนบ้านใกล้ที่สุด-เค [16] พร้อมกำหนดป้ายกำกับข้อมูลด้วยเทคนิค Label Encoding

2) การคัดเลือกคุณลักษณะ

คัดเลือกคุณลักษณะที่เหมาะสมด้วยเทคนิคการคัดกรองข้อมูล (Filter Method) โดยประยุกต์ใช้การวิเคราะห์ความ

แปรปรวนทางเดียว (One-Way ANOVA) เพื่อประเมินความแตกต่างของค่าเฉลี่ยของตัวแปรอิสระระหว่างกลุ่มของมรดกแต่ละแหล่ง โดยคัดเลือกเฉพาะตัวแปรที่มีความแตกต่างอย่างมีนัยสำคัญทางสถิติ ($p\text{-value} < .05$) ควบคู่กับการพิจารณาค่าสถิติ F (F-score) เพื่อจัดลำดับความสำคัญของคุณลักษณะ ทั้งนี้ตัวแปรที่มีค่า F-score สูงแสดงถึงความสามารถในการจำแนกค่าเฉลี่ยของตัวแปรอิสระระหว่างกลุ่มได้ดีกว่า จึงมีแนวโน้มเป็นคุณลักษณะที่มีความสำคัญต่อการสร้างแบบจำลองการจำแนกประเภท จากนั้นทำการวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน (Pearson correlation coefficient) ผ่านแผนภาพความร้อน (Correlation heatmap) เพื่อตรวจสอบปัญหาความสัมพันธ์พหุร่วม (Multicollinearity) การที่ตัวแปรอิสระมีความสัมพันธ์เชิงเส้นระหว่างกันในระดับที่สูงเกินไป ซึ่งอาจส่งผลกระทบต่อความแม่นยำของแบบจำลอง

3) การแบ่งชุดข้อมูล

ชุดข้อมูลทั้งหมด ถูกแบ่งออกเป็นชุดข้อมูลฝึกสอน (Training Set) และชุดข้อมูลทดสอบ (Test Set) ในอัตราส่วน 80:20 โดยชุดข้อมูลฝึกสอนจะถูกนำไปใช้สำหรับการสร้างแบบจำลอง และปรับแต่งค่าไฮเปอร์พารามิเตอร์ ในขณะที่ชุดข้อมูลทดสอบจะนำไปใช้ประเมินประสิทธิภาพขั้นสุดท้ายของแบบจำลอง

3.2.4 การสร้างแบบจำลอง

นำชุดข้อมูลฝึกสอนมาสร้างแบบจำลองการเรียนรู้ของเครื่องเพื่อการจำแนกแหล่งกำเนิดของมรดก 6 ประเภท ดังนี้

1) การวิเคราะห์จำแนกเชิงเส้น (Linear Discriminant Analysis: LDA) เป็นวิธีการเรียนรู้แบบมีผู้สอน (Supervised Learning) ที่ใช้สำหรับการจำแนกข้อมูลตั้งแต่ 2 กลุ่มขึ้นไป เพื่อค้นหาการรวมเชิงเส้นของตัวแปรที่สามารถเพิ่มความแตกต่างระหว่างกลุ่มให้มากที่สุด และลดความแปรปรวนภายในกลุ่มให้ต่ำที่สุด ผ่านการวิเคราะห์เมทริกซ์ความแปรปรวนระหว่างกลุ่ม (Between-class Scatter Matrix) และเมทริกซ์ความแปรปรวนภายในกลุ่ม (Within-class Scatter Matrix) จากนั้นจึงแก้ปัญหา Eigenvalue เพื่อหาแกนจำแนกเชิงเส้น (Linear Discriminants) ที่เหมาะสม โดยค่า Eigenvalue ใช้สะท้อนความสามารถในการจำแนกของแต่ละแกนจำแนก ซึ่งค่า

Eigenvalue ที่สูงกว่าจะสามารถอธิบายความแตกต่างระหว่างกลุ่มได้ดีกว่า [17]

2) การถดถอยโลจิสติกส์ (Logistic Regression: LR) เป็นแบบจำลองเชิงเส้นสำหรับการจำแนกประเภท โดยสร้างสมการเชิงเส้นจากตัวแปรอิสระเพื่อประมาณความน่าจะเป็นของการเป็นสมาชิกในแต่ละกลุ่ม สำหรับปัญหาการจำแนกหลายกลุ่ม งานวิจัยนี้ใช้ Multinomial Logistic Regression ซึ่งประยุกต์ใช้ฟังก์ชัน Softmax ในการแปลงค่าผลรวมเชิงเส้นให้เป็นค่าความน่าจะเป็นของแต่ละกลุ่ม โดยผลรวมของความน่าจะเป็นของทุกกลุ่มมีค่าเท่ากับ 1 และกำหนดผลการจำแนกตามกลุ่มที่มีค่าความน่าจะเป็นสูงสุด [18]

3) นาอิวเบย์ (Naïve Bayes) เป็นอัลกอริทึมสำหรับการจำแนกประเภทที่อาศัยทฤษฎีบทของเบย์ (Bayes' theorem) โดยคำนวณความน่าจะเป็นแบบมีเงื่อนไขของการเป็นสมาชิกในแต่ละกลุ่มจากค่าของคุณลักษณะ ภายใต้สมมติฐานว่าคุณลักษณะทุกตัวเป็นอิสระต่อกันแบบมีเงื่อนไข (Conditional Independence) [19] เนื่องจากข้อมูลที่ใช้ในงานวิจัยนี้เป็นข้อมูลเชิงปริมาณ จึงเลือกใช้ Gaussian Naïve Bayes (GNB) ซึ่งสมมติว่าค่าของคุณลักษณะแต่ละตัวภายในแต่ละกลุ่มมีการแจกแจงแบบปกติ (Gaussian Distribution)

4) เพื่อนบ้านใกล้ที่สุด-เค (k-Nearest Neighbors) เป็นอัลกอริทึมแบบไม่อิงพารามิเตอร์ (Non-parametric Algorithm) [20] ที่ใช้หลักการวัดระยะห่างระหว่างข้อมูลเพื่อประเมินความใกล้เคียง เมื่อมีข้อมูลใหม่เข้าสู่แบบจำลอง จะคำนวณระยะห่างระหว่างข้อมูลดังกล่าวกับข้อมูลทั้งหมดในชุดฝึก แล้วเลือกข้อมูลที่อยู่ใกล้ที่สุดจำนวน k ตัวอย่าง (Nearest Neighbors) เพื่อใช้ในการพิจารณาจำแนกประเภท โดยกำหนดผลการจำแนกเป็นกลุ่มที่มีจำนวนสมาชิกมากที่สุดในบรรดาเพื่อนบ้านทั้ง k ตัวอย่าง

5) ป่าสุ่ม (Random Forest) เป็นอัลกอริทึมการเรียนรู้แบบกลุ่ม (Ensemble Learning) ที่สร้างแบบจำลองการจำแนกประเภทจากการรวมผลลัพธ์ของต้นไม้ตัดสินใจ (Decision Trees) หลายต้น แต่ละต้นถูกฝึกด้วยชุดข้อมูลย่อยที่ได้จากการสุ่มตัวอย่างแบบ Bootstrap (Bagging) และสุ่มเลือกบางส่วนของคุณลักษณะในแต่ละโหนดเพื่อใช้ในการแบ่งข้อมูล การแบ่งโหนดจะเลือกเงื่อนไขที่ทำให้ค่าความไม่บริสุทธิ์

(Impurity) ลดลงมากที่สุด ซึ่งสามารถวัดได้ด้วย Gini Impurity หรือ Entropy จนกว่าจะถึงเกณฑ์หยุดที่กำหนด ผลการจำแนกประเภทสุดท้ายได้จากการโหวตเสียงข้างมาก (Majority Voting) ของต้นไม้ตัดสินใจทั้งหมด [19]

6) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) เป็นอัลกอริทึมสำหรับการจำแนกประเภทที่มุ่งค้นหาเส้นไฮเปอร์เพลน (Hyperplane) ที่สามารถแบ่งแยกกลุ่มข้อมูลได้ดีที่สุด โดยทำให้ระยะขอบ (Margin) ระหว่างกลุ่มข้อมูลมีค่าสูงสุด (Maximum Margin) ข้อมูลที่อยู่ใกล้กับเส้นไฮเปอร์เพลนมากที่สุดเรียกว่า ซัพพอร์ตเวกเตอร์ (Support Vectors) ซึ่งเป็นข้อมูลสำคัญที่ใช้กำหนดตำแหน่งของเส้นแบ่งกลุ่ม ในกรณีที่ข้อมูลไม่สามารถแบ่งแยกเชิงเส้นได้ จะประยุกต์ใช้ฟังก์ชันเคอร์เนล (Kernel Function) เช่น Radial Basis Function (RBF) Kernel เพื่อคำนวณข้อมูลในปริภูมิคุณลักษณะที่มีมิติสูงขึ้น ทำให้สามารถแบ่งแยกกลุ่มข้อมูลได้อย่างมีประสิทธิภาพ [22]

สำหรับการปรับแต่งพารามิเตอร์ (Hyperparameter Tuning) งานวิจัยนี้ใช้เทคนิค Grid Search ร่วมกับการตรวจสอบความถูกต้องแบบไขว้ชนิด 5-fold cross-validation เพื่อค้นหาชุดพารามิเตอร์ที่ให้ค่าความแม่นยำ (Accuracy) สูงที่สุดสำหรับแต่ละอัลกอริทึมการเรียนรู้ของเครื่อง โดยกำหนดค่า $k=5$ เนื่องจากเป็นค่าที่ให้ความสมดุลระหว่างความน่าเชื่อถือในการประเมินประสิทธิภาพของแบบจำลองและการะในการคำนวณ [23]

4.2.5 การประเมินผล

ประเมินประสิทธิภาพของแบบจำลองในการจำแนกแหล่งกำเนิดของมรดก โดยนำไปทดสอบกับชุดข้อมูลทดสอบ (Testing set) เครื่องมือหลักที่ใช้ในการวัดประสิทธิภาพของการจำแนกประเภทแบบหลายกลุ่ม (Multiclass Classification) คือ เมทริกซ์ความสับสน (Confusion Matrix) ซึ่งจะแสดงผลการเปรียบเทียบระหว่างค่าที่แบบจำลองทำนายได้กับค่าจริงในแต่ละกลุ่มเป้าหมาย การวัดผลจะใช้วิธี One-vs-Rest (OVR) เพื่อพิจารณาความสนใจเทียบกับกลุ่มอื่นๆ ทั้งหมด โดยทำการพิจารณาจาก 4 ค่า ได้แก่ ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าความถ่วงดุล (F1 Score) [21] สามารถคำนวณได้จากสมการที่ (1) – (5)

1) ค่าความถูกต้อง (Accuracy) เป็นตัววัดที่ใช้ประเมินสัดส่วนของจำนวนตัวอย่างที่แบบจำลองจำแนกได้ถูกต้องต่อจำนวนตัวอย่างทั้งหมด โดยพิจารณาผลการจำแนกของทุกกลุ่ม หากค่า Accuracy เข้าใกล้ร้อยละ 100 แสดงว่าแบบจำลองมีประสิทธิภาพในการจำแนกข้อมูลสูง

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \quad (1)$$

2) ค่าความแม่นยำ (Precision)

$$Precision = \frac{TP}{(TP + FP)} \quad (2)$$

3) ค่าความระลึก (Recall)

$$Recall = \frac{TP}{(TP + FN)} \quad (3)$$

4) ค่าประสิทธิภาพโดยรวม (F1 Score)

$$F1 = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \quad (4)$$

โดยที่

True Positive (TP) คือ จำนวนข้อมูลที่เป็นของกลุ่มที่พิจารณา และแบบจำลองสามารถจำแนกได้ถูกต้องว่าเป็นกลุ่มนั้น

True Negative (TN) คือ จำนวนข้อมูลที่ไม่ใช่กลุ่มที่พิจารณา และแบบจำลองสามารถจำแนกได้ถูกต้องว่าไม่ใช่กลุ่มนั้น

False Positive (FP) คือ จำนวนข้อมูลที่ไม่ใช่กลุ่มที่พิจารณา แต่แบบจำลองจำแนกผิดว่าเป็นกลุ่มนั้น

False Negative (FN) คือ จำนวนข้อมูลที่เป็นของกลุ่มที่พิจารณา แต่แบบจำลองจำแนกผิดว่าเป็นกลุ่มอื่น

3.2.6 การนำไปใช้งาน

แบบจำลองที่ได้รับการพัฒนาสามารถเป็นเครื่องมือสนับสนุนในการจำแนกแหล่งกำเนิดของมรกตได้อย่างถูกต้องและแม่นยำมากขึ้น

4. ผลการวิจัย

4.1 ผลการจัดการกับข้อมูล

จากการตรวจสอบข้อมูลพบว่ามีจำนวนข้อมูลทั้งหมด 1,088 แถว และ 18 คอลัมน์ (รูปที่ 2) โดยคอลัมน์ Origin คือ ตัวแปรตาม (Dependent Variables: X) ประกอบด้วยแหล่งมรกตจำนวน 4 แหล่ง ได้แก่ Afghanistan (116 ตัวอย่าง) Colombia (306 ตัวอย่าง) Zambia (Kafubu) (393 ตัวอย่าง) และ Zambia (Musakashi) (273 ตัวอย่าง) ไม่พบข้อมูลสูญหาย ส่วนตัวแปรต้น (Independent Variables: y) คือ ค่าความเข้มข้นของธาตุร่องรอย จำนวน 16 ชนิด ซึ่งเป็นข้อมูลเชิงปริมาณ ได้แก่ Li, Na, Mg, K, Ca, Sc, V, Cr, Mn, Fe, Co, Ni, Zn, Ga, Rb และ Cs แสดงในหน่วย Parts per million by weight (ppmw) อย่างไรก็ตามในบางตัวอย่างแสดงค่าความเข้มข้นของธาตุเป็น “b.d.” (Below Detection Limit) ซึ่งเกิดจากความเข้มข้นของธาตุต่ำกว่าขีดจำกัดการตรวจวัดของเครื่องมือ (Limit of Detection: LOD) จากการตรวจสอบข้อมูลพบเฉพาะในธาตุ K, Ca, Co, Ni และ Zn และเนื่องจากไม่ทราบค่า LOD ของแต่ละธาตุอย่างชัดเจนจึงไม่สามารถประมาณค่าทดแทนด้วยวิธีเชิงสถิติ เช่น การแทนด้วยค่ากึ่งหนึ่งของ LOD (LOD/2) ได้อย่างเหมาะสม ดังนั้นงานวิจัยนี้จึงเลือกแทนค่าดังกล่าวด้วยวิธีแทนค่าด้วยหลายค่า (Multiple Imputation; MI) เพื่อให้สามารถนำข้อมูลเข้าสู่กระบวนการวิเคราะห์ได้ นอกจากนี้ได้ทำการลบข้อมูลในคอลัมน์ Analysis No. ออกไปเนื่องจากไม่มีผลต่อการวิเคราะห์

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1088 entries, 0 to 1087
Data columns (total 18 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Analysis no.          1088 non-null   object
1   Origin                1088 non-null   object
2   Li                    1088 non-null   float64
3   Na                    1088 non-null   float64
4   Mg                    1088 non-null   float64
5   K                     1082 non-null   float64
6   Ca                    1068 non-null   float64
7   Sc                    1088 non-null   float64
8   V                     1088 non-null   float64
9   Cr                    1088 non-null   float64
10  Mn                    1088 non-null   float64
11  Fe                    1088 non-null   float64
12  Co                    617 non-null    float64
13  Ni                    1033 non-null   float64
14  Zn                    976 non-null    float64
15  Ga                    1088 non-null   float64
16  Rb                    1088 non-null   float64
17  Cs                    1088 non-null   float64
dtypes: float64(16), object(2)
memory usage: 153.1+ KB
```

รูปที่ 2. แสดงรายละเอียดข้อมูลที่ใช้ในงานวิจัยนี้

นอกจากนี้ได้ทำการสังเคราะห์ตัวอย่างเพิ่มในกลุ่มที่มีจำนวนตัวอย่างน้อย เฉพาะในชุดข้อมูลฝึกสอน (Training Set) จำนวนข้อมูลก่อนและหลังการปรับสมดุลด้วยวิธีการสังเคราะห์ข้อมูลเพิ่ม (SMOTE) ทำให้ตัวอย่างในแต่ละกลุ่มมีจำนวนเท่ากันที่ 314 ตัวอย่าง (ตารางที่ 1)

ตารางที่ 1. แสดงผลการปรับสมดุลจำนวนข้อมูลด้วยวิธี SMOTE

กลุ่ม	จำนวนตัวอย่าง ก่อนทำ SMOTE	จำนวนตัวอย่าง หลังทำ SMOTE
Afghanistan	93	314
Colombia	245	314
Zambia (Kafubu)	314	314
Zambia (Musakashi)	218	314

สำหรับผลลัพธ์การติดป้ายกำกับข้อมูลด้วยเทคนิค Label Encoding แสดงในตารางที่ 2

ตารางที่ 2. แสดงผลการติดป้ายกำกับให้กับกลุ่มตัวอย่าง

กลุ่มตัวอย่าง	ป้ายกำกับ
Afghanistan	0
Colombia	1
Zambia (Kafubu)	2
Zambia (Musakashi)	3

4.2 ผลการคัดเลือกคุณลักษณะ

ทำการคัดเลือกคุณลักษณะสำหรับสร้างแบบจำลองด้วยเทคนิค One-Way ANOVA พบว่าค่าสถิติ F (F-Score) ของตัวแปรต้นอยู่ระหว่าง 32.665 ถึง 2270.786 (ตารางที่ 3) ในงานวิจัยนี้ทดลองคัดเลือกคุณลักษณะ 11 อันดับแรกที่มีค่าสถิติ F สูงกว่า 300 ซึ่งแสดงค่าเฉลี่ยของกลุ่มแตกต่างกันสูง มีแนวโน้มเพิ่มประสิทธิภาพในการจำแนกกลุ่มได้ดี ได้แก่ ลิเทียม (Li) โซเดียม (Na) เหล็ก (Fe) แมกนีเซียม (Mg) ซีเซียม (Cs) แมงกานีส (Mn) รูบิเดียม (Rb) นิกเกิล (Ni) วานาเดียม (V) สังกะสี (Zn) และแคลเซียม (Ca) โดยมีค่า F-score อยู่ระหว่าง 305.690 ถึง 2270.786 นอกจากนี้ตัวแปรดังกล่าวยังแสดงค่า p-value น้อยกว่า 0.05 ซึ่งแสดงให้เห็นว่ามีความสามารถในการจำแนกกลุ่มได้อย่างมีนัยสำคัญทางสถิติ

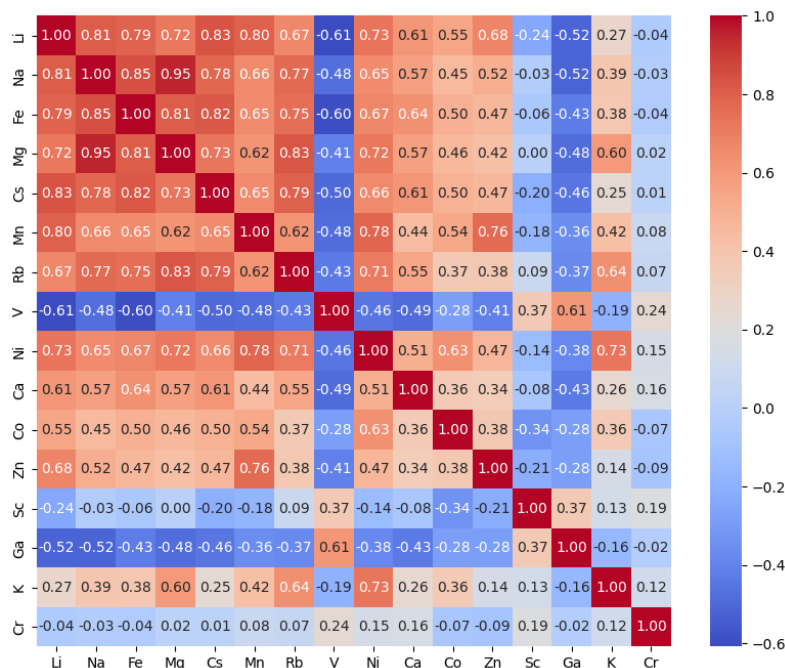
ตารางที่ 3. แสดงค่าสถิติ F และระดับนัยสำคัญของธาตุแต่ละชนิดจากการวิเคราะห์ด้วยเทคนิค One-Way ANOVA

ตัวแปรต้น	คำอธิบาย	F-score	p-value
Li	ลิเทียม	2270.786	< 0.001
Na	โซเดียม	1762.535	< 0.001
Fe	เหล็ก	1260.116	< 0.001
Mg	แมกนีเซียม	797.837	< 0.001
Cs	ซีเซียม	700.895	< 0.001
Mn	แมงกานีส	484.035	< 0.001
Rb	รูบิเดียม	383.838	< 0.001
Ni	นิกเกิล	339.728	< 0.001
V	วานาเดียม	336.771	< 0.001
Zn	สังกะสี	309.685	< 0.001
Ca	แคลเซียม	305.690	< 0.001
Co	โคบอลต์	223.385	< 0.001
Sc	สแกนเดียม	206.622	< 0.001
Ga	แกเลียม	163.878	< 0.001

K	โพแทสเซียม	56.902	< 0.001
Cr	โครเมียม	32.665	< 0.001

ในการประเมินความสัมพันธ์ระหว่างตัวแปรและลดความซ้ำซ้อนของข้อมูล ผู้วิจัยได้ทำการวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน (Pearson's correlation coefficient: r) ซึ่งค่าที่ได้จะอยู่ในช่วงระหว่าง -1 ถึง 1 จากผลการวิเคราะห์พบว่า โซเดียม (Na) และแมกนีเซียม (Mg) มี

ความสัมพันธ์เชิงบวกในระดับสูงมาก โดยมีค่า r เท่ากับ 0.95 (รูปที่ 3) กรณีนี้อาจก่อให้เกิดปัญหา Multicollinearity ในกระบวนการสร้างแบบจำลองเพื่อหลีกเลี่ยงปัญหาดังกล่าว จึงได้พิจารณาค่าสถิติ F ประกอบการตัดสินใจ พบว่า Na มีค่าสถิติ F สูงกว่า Mg จึงทำการตัดตัวแปร Mg ออกจากชุดข้อมูล ดังนั้นคุณลักษณะที่เหมาะสมสำหรับใช้เป็นข้อมูลนำเข้าในการพัฒนาแบบจำลองในงานวิจัยนี้ ประกอบด้วย 10 ธาตุ ได้แก่ Li, Na, Fe, Co, Cs, Mn, Rb, Ni, V และ Ca



รูปที่ 3. แผนภาพความร้อนแสดงค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน (Pearson Correlation Heatmap) ของคุณลักษณะในชุดข้อมูลการวิจัย

4.3 ผลการวัดประสิทธิภาพแบบจำลอง

แบบจำลอง	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
k-NN	95.87	95.94	95.87	95.87
NB	99.08	99.16	99.08	99.10
SVM	86.70	86.82	86.70	86.73
LR	86.24	86.43	86.24	86.24
LDA	84.86	85.60	84.86	84.98
RF	99.54	99.55	99.54	99.54

การวัดประสิทธิภาพของแบบจำลองทั้ง 6 ประเภท ด้วยเครื่องมือ Confusion matrix แสดงในตารางที่ 4 การวัดผลจะอ้างอิงค่าพารามิเตอร์ที่เหมาะสมที่สุด โดยแสดงผลการประเมินด้วยค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าความถ่วงดุล (F-1 Score) โดยวิธี Weighted Averaging

ตารางที่ 4. ผลการวัดประสิทธิภาพของแบบจำลองทั้ง 6 อัลกอริทึม

จากผลการศึกษาพบว่าแบบจำลองป่าสุ่ม (Random Forest) ให้ประสิทธิภาพโดยรวมสูงสุด สามารถจำแนกแหล่งกำเนิดมลพิษได้อย่างถูกต้อง โดยมีค่าความถูกต้อง ร้อยละ 99.54 ค่าความแม่นยำ ร้อยละ 99.55 ค่าความระลึก ร้อยละ 99.54 และค่าความถ่วงดุล ร้อยละ 99.54 รองลงมา คือแบบจำลองเพื่อนบ้านใกล้ที่สุด-เค (k-Nearest Neighbor) ให้ค่าความค่าความถูกต้อง ร้อยละ 95.87 ค่าความแม่นยำ ร้อยละ 95.94 ค่าความระลึก ร้อยละ 95.87 และค่าความถ่วงดุล ร้อยละ 95.87 ในขณะที่แบบจำลองที่มีประสิทธิภาพต่ำสุดในการศึกษานี้ คือ แบบจำลองการวิเคราะห์จำแนกประเภทเชิงเส้น (Linear Discriminant Analysis) ซึ่งให้ค่าความค่าความถูกต้อง ร้อยละ 84.86 ค่าความแม่นยำ ร้อยละ 85.60 ค่าความระลึก ร้อยละ 84.86 และค่าความถ่วงดุล ร้อยละ 84.98

นอกจากนี้ เพื่อยืนยันประสิทธิภาพของแบบจำลองป่าสุ่ม ได้มีการตรวจสอบด้วยวิธี 5-Fold Cross Validation พบว่าค่าความถูกต้องในแต่ละรอบดังแสดงในตารางที่ 5 โดยมีค่าเฉลี่ยความถูกต้องเท่ากับร้อยละ 99.72 แสดงให้เห็นว่าแบบจำลองมีความสม่ำเสมอในการเรียนรู้และสามารถทำการจำแนกข้อมูลที่ไม่เคยเห็นมาก่อนได้อย่างมีประสิทธิภาพ

ตารางที่ 5. ผลการวัดประสิทธิภาพของแบบจำลองป่าสุ่มด้วยเทคนิค 5-Fold Cross Validation

รอบที่	ค่าความถูกต้อง
1	100%
2	100%
3	100%
4	98.62%
5	100%
ค่าเฉลี่ยความถูกต้อง	99.72
ค่าเบี่ยงเบนมาตรฐาน (S.D.)	0.55

5. บทสรุปและการอภิปราย

งานวิจัยนี้มุ่งเน้นไปที่การพัฒนาแบบจำลองการเรียนรู้ของเครื่องเพื่อจำแนกแหล่งกำเนิดของมลพิษจำนวน 4 แหล่งที่สำคัญ โดยผ่านกระบวนการเลือกคุณลักษณะเฉพาะตัวแปรที่สามารถจำแนกกลุ่มได้ดีในการสร้างแบบจำลองเรียนรู้ และส่งผลให้แบบจำลองการเรียนรู้มีประสิทธิภาพสูงในการจำแนก จากการวิเคราะห์ด้วยเทคนิค One-Way Anova และสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน พบว่าคุณลักษณะที่สามารถจำแนกกลุ่มได้ดีมีอยู่จำนวน 10 ธาตุ ได้แก่ Li, Na, Fe, Co, Cs, Mn, Rb, Ni, V และ Ca ซึ่งสอดคล้องกับงานวิจัยของ Alonso-Perez et al. [25] ซึ่งทำการคัดเลือกธาตุร่องรอยจำนวน 8 ธาตุ ได้แก่ Li, V, Cr, Fe, Sc, Ga, Rb และ Cs นำมาสร้างแบบจำลองการเรียนรู้ของเครื่องเพื่อจำแนกมลพิษจากแหล่งต่าง ๆ ได้อย่างมีประสิทธิภาพ นอกจากนี้ยังสอดคล้องกับการศึกษาของ มณฑิรา เสนิงศ์ ณ อุทยา และคณะ [21] คัดเลือกธาตุร่องรอย จำนวน 11 ธาตุ ได้แก่ Li, Na, Mg, K, Sc, V, Cr, Fe, Ga, Rb และ Cs ในการสร้างแบบจำลองการเรียนรู้ของเครื่องที่สามารถจำแนกแหล่งกำเนิดของมลพิษได้อย่างมีประสิทธิภาพเช่นเดียวกัน จากงานวิจัยนี้สามารถสรุปได้ว่าคุณลักษณะจำนวน 5 ธาตุ ได้แก่ Li, V, Fe, Rb และ Cs เป็นกลุ่มธาตุร่องรอยที่มีประสิทธิภาพสูงในการจำแนกแหล่งกำเนิดของมลพิษ ซึ่งสอดคล้องกับงานวิจัยนี้โดยพบว่าธาตุดังกล่าวมีค่าสถิติ F อยู่ใน 10 อันดับแรก ทั้งนี้ Alonso-Perez et al. [25] อธิบายเพิ่มเติมว่ามลพิษจากแหล่งโคลอมเบีย มีความเข้มข้นของธาตุรูบิเดียมในระดับต่ำเพียง 0.2–3.3 ppmw เมื่อเทียบกับแหล่งอื่น ๆ เช่น แหล่งมาดากัสการ์ (23–114 ppmw) และแหล่งแซมเบีย (16–138 ppmw) ดังนั้นธาตุรูบิเดียมจึงเป็นตัวแปรที่สำคัญในการจำแนกมลพิษจากโคลอมเบียออกจากแหล่งอื่น เช่นเดียวกับการศึกษาของ Saeseaw et al. [3] ที่สนับสนุนว่ามลพิษจากโคลอมเบียประกอบด้วยธาตุอัลคาไลที่มีปริมาณความเข้มข้นต่ำกว่าแหล่งอื่น โดยเฉพาะธาตุ Li, K, Rb, Cs

และ Fe ในขณะที่มรดกจากแหล่งกาฟูบูในประเทศแซมเบียมีปริมาณธาตุ Cs, Fe และ Li ก่อนข้างสูงกว่าแหล่งอื่น ๆ

การวิจัยนี้แสดงให้เห็นว่าแบบจำลองป่าสุ่ม (Random Forest) ให้ประสิทธิภาพสูงสุดในการจำแนกกำเนิดของมรดก โดยให้ค่าความถูกต้องที่ร้อยละ 99.54 รองลงมาคือแบบจำลองนาอิวเบย์ ซึ่งยังถือว่ามีประสิทธิภาพสูงเช่นกัน โดยมีค่าความถูกต้องร้อยละ 99.08 ซึ่งผลการวิจัยดังกล่าวนี้สอดคล้องกับการศึกษาของ มณฑิรา เสนิงค์ ณ อุรุยา และคณะ [26] ซึ่งได้ศึกษาและพัฒนาแบบจำลองการเรียนรู้ของเครื่องในการจำแนกแหล่งกำเนิดของมรดกจาก 5 แหล่ง ได้แก่ ประเทศโคลอมเบีย แซมเบีย เอธิโอเปีย ปากีสถาน และอัฟกานิสถาน ซึ่งผลการวิจัยพบว่าแบบจำลองป่าสุ่ม และโครงข่ายประสาทเทียม (Artificial Neural Network) โดยให้ค่าความถูกต้องสูงสุดเท่ากันที่ร้อยละ 97 ในขณะที่แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนให้ประสิทธิภาพต่ำสุดที่ร้อยละ 86.70 นอกจากนี้ในงานวิจัยนี้ ยังสอดคล้องกับการศึกษาของ Alonso-Perez et al. [25] ซึ่งทำการวัดประสิทธิภาพของแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนเพื่อจำแนกมรดก 13 แหล่ง ซึ่งพบว่าแบบจำลองให้ค่าความถูกต้องเพียงร้อยละ 79 โดยสาเหตุเกิดจากความคล้ายคลึงกันของข้อมูลระหว่างคลาส (Inter-Class Similarities) ซึ่งส่งผลให้เส้นแบ่งระหว่างกลุ่มมีความซับซ้อนและยากต่อการแยกอย่างชัดเจน

จากผลการวิจัยนี้สามารถนำไปแก้ปัญหาที่เกิดขึ้นบ่อยครั้งในการจำแนกแหล่งกำเนิดของมรดก ซึ่งบ่อยครั้งนักอณูเคมีจะใช้วิธีการพล็อตกราฟเปรียบเทียบระหว่างธาตุสองธาตุ (Bivariate Plot) เช่น การศึกษาของ Krzemnicki et al. [15] ทำการพล็อตกราฟเปรียบเทียบระหว่างธาตุรูบิเดียม (Rb) และซีเซียม (Cs) เพื่อจำแนกมรดกจากประเทศโคลอมเบีย และแซมเบีย (มุซาคาซิ) แต่กลับพบว่าข้อมูลเคมีของทั้งสองแหล่งยังคงทับซ้อนกันอยู่ เช่นเดียวกับการศึกษาของ Saeseaw et al. [3] พบว่าการพล็อตกราฟเทียบกันระหว่าง

โพแทสเซียม (K) และเหล็ก (Fe) ไม่สามารถแยกแหล่งมรดกจากประเทศโคลอมเบีย จีน และแซมเบีย (มุซาคาซิ) ออกจากกันได้ชัดเจน ทั้งนี้จำเป็นต้องใช้ข้อมูลทางสเปกโทรสโกปีและลักษณะเฉพาะของมลทินภายในประกอบการตัดสินใจเพิ่มเติม

6. ข้อเสนอแนะสำหรับการวิจัยในอนาคต

ในการวิจัยครั้งนี้ แม้ว่าแบบจำลองป่าสุ่มจะให้ค่าความถูกต้องสูงถึงร้อยละ 99.54 แต่เมื่อพิจารณาแล้ว อาจเกิดจากจำนวนข้อมูลที่มีจำกัด ซึ่งส่งผลให้แบบจำลองเรียนรู้ข้อมูลที่น้อยเกินไปจนทำให้ง่ายต่อการจำแนก ดังนั้นแนวทางการพัฒนาในอนาคตควรปรับปรุงข้อมูลให้มีความซับซ้อนและความท้าทายมากยิ่งขึ้น ดังนี้

- 1) เพิ่มปริมาณข้อมูลสำหรับชุดข้อมูลฝึกสอน และชุดข้อมูลทดสอบ รวมถึงการจัดการจำนวนตัวอย่างในแต่ละคลาสให้มีความใกล้เคียงกัน เพื่อพัฒนาแบบจำลองการเรียนรู้ของเครื่องให้มีความน่าเชื่อถือมากยิ่งขึ้น
- 2) เพิ่มความหลากหลายของข้อมูลโดยรวบรวมมรดกจากแหล่งอื่น ๆ ซึ่งในปัจจุบันพบว่ามรดกอีกหลายแหล่งมากที่มีจำหน่ายในตลาดอณูเคมี เช่น บราซิล มาดากัสการ์ เอธิโอเปีย รัสเซีย และปากีสถาน เพื่อเพิ่มความถูกต้องและแม่นยำของแบบจำลองในกรณีที่เจอมรดกจากแหล่งอื่น ๆ
- 3) บูรณาการร่วมกับข้อมูลทางอณูเคมีวิทยาเพื่อเพิ่มขีดความสามารถให้กับแบบจำลอง เช่น ลักษณะเด่นของมลทินภายใน ตำแหน่งของสเปกตรัมที่ได้จากเทคนิค UV-Vis-NIR / Raman / Photoluminescence Spectroscopy เป็นต้น
- 4) ประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่องที่หลากหลายมากยิ่งขึ้น เช่น การเรียนรู้เชิงลึก (Deep Learning) หรือการเรียนรู้แบบผสม (Ensemble Learning) เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่องขั้นสูงต่อไป

5) พัฒนาแบบจำลองการเรียนรู้ของเครื่องสำหรับจำแนกแหล่งกำเนิดของอัญมณีชนิดอื่น ๆ ที่มีความสำคัญในเชิงพาณิชย์ เช่น ทับทิม ไพลิน สปิเนล และทัวร์มาลีน เป็นต้น

เอกสารอ้างอิง

[1] S. F. McClure, T. M. Moses, and J. E. Shigley, "The geographic origin dilemma," *Gems & Gemology*, vol. 55, no. 4, pp. 450–453, 2019.

[2] J.-L. Archuleta, "The color of responsibility: Ethical issues and solutions in colored gemstones," *Gems & Gemology*, vol. 52, no. 2, pp. 114–131, 2016.

[3] S. Saeseaw, N. D. Renfro, A. C. Palke, Z. Sun, and S. F. McClure, "Geographic origin determination of emerald," *Gems & Gemology*, vol. 55, no. 4, pp. 478–507, 2019.

[4] D. L. Wood and K. Nassau, "The characterization of beryl and emerald by visible and infrared absorption spectroscopy," *American Mineralogist*, vol. 53, pp. 777–800, 1968.

[5] E. Fritsch and G. R. Rossman, "An update on color in gems. Part I: Introduction and colors caused by dispersed metal ions," *Gems & Gemology*, vol. 23, pp. 126–139, 1987.

[6] D. Schwarz, G. Giuliani, G. Grundmann, and M. Glas, "The origin of emerald...a controversial topic," *extraLapis English*, no. 2, pp. 18–21, 2002.

[7] M. Awad and R. Khanna, *Machine Learning in Action: Examples*. New York, NY, USA: Apress, 2015.

[8] T. Blodgett and A. H. Shen, "Application of discriminant analysis in gemology: Country-of-origin separation in colored stones and distinguishing HPHT-treated diamonds," *Gems & Gemology*, vol. 47, no. 2, p. 145, 2011.

[9] Z. Luo, M. Yang, and A. H. Shen, "Origin determination of dolomite-related white nephrite through IB-LDA," *Gems & Gemology*, vol. 51, no. 3, pp. 268–278, 2015.

[10] A. Homkrajac, Z. Sun, and S. C. Shih, "Gemological and chemical characteristics of natural freshwater pearls from the

Mississippi River system," *Gems & Gemology*, vol. 55, no. 2, pp. 164–185, 2019.

[11] H. A. O. Wang and M. S. Krzemnicki, "Multi-element analysis of minerals using laser ablation inductively coupled plasma time of flight mass spectrometry and geochemical data visualization using t-distributed stochastic neighbor embedding: Case study on emeralds," *Journal of Analytical Atomic Spectrometry*, vol. 36, no. 3, pp. 518–527, 2021.

[12] Y. Zhong, A. H. Shen, Z. Zhang, M. Ye, and Y. Han, "Application of machine learning algorithms in the geographical origin determination of peridot," *Journal of Gems & Gemology*, vol. 25, no. 6, pp. 65–75, 2023 (in Chinese).

[13] M. Seneewong-Na-Ayutthaya, B. Wanthanachaisaeng, W. Suwanmanee, T. Lhuamporn, P. Kamkeaw, S. Phibanchon, T. Sripoonjan, and T. Leelawatanasuk, "An implementation of machine learning in ruby and sapphire origin determination," in *Proc. 37th International Gemmological Conference (IGC 2023)*, Tokyo, Japan, 2023.

[14] G. Mariscal, Ó. Marbán, and C. Fernández, "A survey of data mining and knowledge discovery process models and methodologies," *The Knowledge Engineering Review*, vol. 25, no. 2, pp. 137–166, 2010.

[15] M. S. Krzemnicki, H. A. O. Wang, M. Wälle, P. Lefèvre, W. Zhou, and L. E. Cartier, "Gemmological characterisation of emeralds from Musakashi, Zambia, and implications for their geographic origin determination," *Journal of Gemmology*, vol. 39, no. 4, pp. 338–350, 2024.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[17] A. Tharwat, T. Gaber, A. Ibrahim, and A. E. Hassanien, "Linear discriminant analysis: A detailed tutorial," *AI Communications*, vol. 30, no. 2, pp. 169–190, 2017.

- [18] C. Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," *Journal of Educational Research*, vol. 96, no. 1, pp. 3–14, 2002.
- [19] P. Domingos and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss," *Machine Learning*, vol. 29, no. 2–3, pp. 103–130, 1997.
- [20] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.
- [21] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.
- [22] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Schölkopf, "Support vector machines," *IEEE Intelligent Systems*, vol. 13, no. 4, pp. 18–28, 1998.
- [23] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York, NY, USA: Springer, 2013.
- [24] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.
- [25] F. Alonso-Pérez, L. T. Grisi, F. G. Sardi, A. Martínez, A. L. de Oliveira, L. Lobo-Guerrero, G. Barone, and S. Raneri, "Geographic origin determination of emeralds using machine learning algorithms and handheld LIBS data," *Minerals*, vol. 14, no. 1, Art. no. 89, 2024.
- [26] มณฑิรา เสนีวงศ์ ณ อยุธยา, ทะนง ลีลาวัฒนสุข, ภูวดล วรธนะชัยแสง, วรดิษฐ์ชนก สุวรรณมณี, ชีวากร สุพรรณ, พิมพ์ภัฏส คำแก้ว, ทศนรา ศรีบุญจั่น, และ อนุรักษ์ บุญมาวงศ์, "กระบวนการพิสูจน์แหล่งกำเนิดอัญมณีโดยใช้เอกลักษณ์ทางเคมีด้วยเทคนิค LA-ICP-MS (เบรล)," ใน *การสัมมนาเผยแพร่งานวิจัยและนวัตกรรมด้านอัญมณีและ*

เครื่องประดับ นูรณาการความร่วมมือหน่วยงานเครือข่ายสู่
การนำไปใช้ประโยชน์, กรุงเทพฯ, ประเทศไทย, 2567.