การจำแนกประเภทของเห็ดระหว่างเห็ดมีพิษหรือไม่มีพิษ โดยใช้เทคนิคการเรียนรู้ของเครื่อง

ภิรมย์พร เถี่ยนมิตรภาพ[1*], วราภรณ์ วิยานนท์[2]

บทคัดย่อ

เห็ดตามธรรมชาติมีทั้งเห็ดที่รับประทานได้และเห็ดที่มีพิษ หากคนรับประทานเห็ดพิษเข้าไปทำให้เกิดอาการเจ็บป่วยและถึงแก่ชีวิตได้ ดังนั้นจึงได้นำเทคนิคการเรียนรู้ของเครื่องมาใช้ในการจำแนกประเภทของเห็ดระหว่างเห็ดพิษและเห็ดที่รับประทานได้ โดยใช้แบบจำลอง 5 แบบจำลอง ได้แก่ Logistic regression, SVM, Decision tree, Random Forest และ XGBoost ร่วมกับการใช้เทคนิคการคัดเลือกและสกัดฟีเจอร์ เพื่อค้นหาฟีเจอร์ที่สำคัญที่สามารถบ่งบอกประเภทของเห็ดได้ โดยผลการศึกษาพบว่าการใช้แบบจำลอง Decision tree ที่ใช้เทคนิค Recursive Feature Elimination (RFE) ในการคัดเลือกฟีเจอร์โดยใช้ training size เท่ากับ 60% และใช้ฟีเจอร์เพียง 3 ฟีเจอร์ได้แก่ กลิ่นของเห็ด (odor) ขนาดครีบเห็ด (gill_size) และสีรอยพิมพ์สปอร์ (spore_print_color) มีความเหมาะสมที่สุดโดยได้ F1 score เท่ากับ 99.32% ซึ่งนำโมเดลที่ได้ไปสร้างต้นแบบเว็บแอปพลิเคชัน

คำสำคัญ : การจำแนกประเภทเห็ด, เทคนิคการเรียนรู้ของเครื่อง, การคัดเลือกฟีเจอร์, การสกัดฟีเจอร์, เว็บแอปพลิเคชัน

---

[1] หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ กรุงเทพฯ 10110

[2] คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ กรุงเทพฯ 10110

* Corresponding author:Tel.: 087-9780903 E-mail address: phiromporn.tia@g.swu.ac.th

Mushroom Classification between Poisonous and Edible Mushrooms using Machine Learning Techniques

Phiromporn Tianmitrapap[1*], Waraporn Viyanon[2]

Abstract

In nature, there are both edible and poisonous mushrooms. Consuming poisonous mushrooms can cause illness and even death. Therefore, machine learning techniques have been employed to classify mushrooms into poisonous and edible types using five models: Logistic Regression, SVM, Decision Tree, Random Forest, and XGBoost. Feature selection and extraction techniques were used to identify important features that can indicate mushroom types. The study found that using the Decision Tree model with Recursive Feature Elimination (RFE) technique and a training size of 60%, with only 3 features: odor, gill size, and spore print color, yielded the best results with an F1 score of 99.32%. The model was then used to build a prototype web application.

Keywords: Mushroom classification, Machine Learning, Feature selection, Feature extraction, Web application.

[1] Data Science, Faculty of Science, Srinakharinwirot University, Bangkok, 10110, Thailand

[2] Faculty of Science, Srinakharinwirot University, Bangkok, 10110, Thailand

* Corresponding author:Tel.: 087-9780903 E-mail address: phiromporn.tia@g.swu.ac.th

1. บทนำ

เห็ดเป็นสิ่งมีชีวิตที่อยู่ในอาณาจักรเห็ดรา (Kingdom fungi) ซึ่งเห็ดจัดเป็นราชั้นสูง ที่อยู่ใน 2 ไฟลัม คือ Ascomycota และ Basidiomycota โดยเห็ดที่คนรู้จักทั่วไปและพบเห็นได้บ่อยจะเป็นเห็ดในไฟลัม Basidiomycota ในอันดับ Agaricales และ Russulales ซึ่งมีชื่อทางวิชาการว่า “เห็ดครีบ (gilled mushroom หรือ agaric mushroom)” [1]

ประเทศไทยตั้งอยู่ในพื้นที่เขตร้อนชื้นและมีฝนตกชุกจึงเหมาะกับการเจริญเติบโตของเห็ด [2] ซึ่งเห็ดตามธรรมชาติมีทั้งเห็ดที่รับประทานได้และเห็ดที่มีพิษ สำหรับเห็ดที่รับประทานได้นั้นให้คุณค่าทางโภชนาการ ส่วนเห็ดพิษคือเห็ดที่สร้างสารพิษชีวภาพทำให้คนเกิดอาการเจ็บป่วยและถึงแก่ชีวิตได้ ดังนั้นการจำแนกประเภทของเห็ดระหว่างเห็ดมีพิษและเห็ดที่รับประทานได้จึงมีความสำคัญ

งานวิจัยที่เกี่ยวกับการจำแนกประเภทเห็ดระหว่างเห็ดพิษกับเห็ดที่รับประทานได้โดยใช้เทคนิคการเรียนรู้ของเครื่อง ซึ่งมี 2 กลุ่มที่สนใจคือ กลุ่มที่ 1: ไม่ได้ทำความสะอาดข้อมูลแต่มีการคัดเลือกหรือสกัดฟีเจอร์ เช่น งานของ Shuhaida Ismail และคณะ [3] ใช้เทคนิค PCA โดยไม่ได้ระบุจำนวนคอมโพแนนท์ และใช้แบบจำลอง Decision tree ซึ่งได้ค่า F1 score = 100% โดยฟีเจอร์ “กลิ่นของเห็ด” เป็นฟีเจอร์ที่มีความสำคัญที่สุด, งานของ S.K. Verma และคณะ [4] มีการสกัดฟีเจอร์แต่ไม่ระบุเทคนิค และมีการแบ่งชุดข้อมูลเป็น 5 อัตราส่วน คือ 40:60, 50:50, 60:40, 70:30 และ 80:20 พบว่าแบบจำลอง ANFIS ที่อัตราส่วน 80:20 มีประสิทธิภาพที่ดีที่สุด (Accuracy = 99.87%) ส่วนกลุ่มที่ 2: มีการทำความสะอาดข้อมูลและคัดเลือกหรือสกัดฟีเจอร์ เช่น งานของ Nadya Chitayae และคณะ [5] มีการลบแถวข้อมูลที่มีค่าว่างจำนวน 2,480 แถว (ในคอลัมน์ชื่อ “stalk_root”) และเลือกฟีเจอร์จำนวน 8 ฟีเจอร์ไปใช้ พบว่าแบบจำลอง Decision tree (CART algorithm) ให้ cross validation score มากที่สุด (91.93%), งานของ V. Vanitha และคณะ [6] ได้ลบฟีเจอร์ “stalk-root” ซึ่งมีค่าว่างจำนวนมาก และคัดเลือกฟีเจอร์ด้วยวิธี wrapper และ filter ได้ฟีเจอร์จำนวน 2 ฟีเจอร์คือ “odor” (กลิ่นของเห็ด) และ “spore_print_color” (สีรอยพิมพ์สปอร์) พบว่าแบบจำลอง Decision tree (J48 algorithm) ให้ F1 score = 99%

ในการศึกษานี้ผู้วิจัยได้สร้างเทคนิคการเรียนรู้ของเครื่องเพื่อช่วยในการจำแนกประเภทของเห็ด โดยใช้ข้อมูลคุณลักษณะของเห็ด ประกอบการใช้เทคนิคการคัดเลือกฟีเจอร์ เพื่อค้นหาฟีเจอร์ที่สำคัญที่สามารถจำแนกประเภทของเห็ดได้

2. วิธีดำเนินการ

งานวิจัยนี้ประกอบด้วยขั้นตอนการดำเนินงานดังรูปที่ 1 โดยอธิบายรายละเอียดในหัวข้อ 2.1-2.6

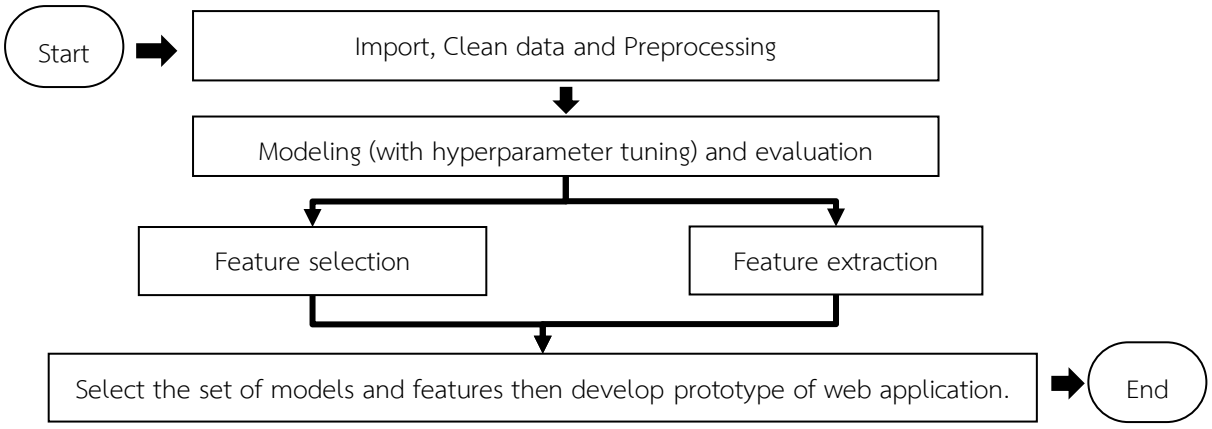

Figure 1 Experimental design

2.1 การนำเข้าชุดข้อมูล ทำความสะอาดข้อมูล และเตรียมข้อมูลเพื่อเข้าสู่แบบจำลอง

งานวิจัยนี้ใช้ชุดข้อมูลสาธารณะของเห็ดครีบจำนวน 23 สายพันธุ์ในตระกูล Agaricus และ Lepiota จำนวน 8,124 แถว ที่มีการระบุว่าเป็นเห็ดพิษหรือเห็ดที่รับประทานได้ โดยได้จากเว็บไซต์ kaggle ซึ่งเป็นข้อมูลที่ Jeff Schlimmer ได้บริจาคให้กับ UCI Machine Learning Repository โดยดึงข้อมูลจากหนังสือ The Audubon Society Field Guide to North American Mushrooms (1981) [7] โดยชุดข้อมูลประกอบด้วยตัวแปร 23 คอลัมน์ ได้แก่ สีหมวกเห็ด (cap_color), ลักษณะพื้นผิวบนหมวกเห็ด (cap_surface), รูปทรงหมวกเห็ด (cap_shape), สีครีบเห็ด (gill_color), ความแนบชิดของครีบกับก้านดอก (gill_attachment), ระยะห่างของครีบเห็ดแต่ละครีบ (gill_spacing), ขนาดครีบเห็ด (gill_size), รูปทรงก้านดอก (stalk_shape), ลักษณะรากที่ปลายก้านดอก (stalk_root), สีและลักษณะของพื้นผิวก้านดอกบริเวณเหนือวงแหวน และบริเวณด้านล่างของวงแหวน (stalk_color_above_ring stalk_color_below_ring, stalk_surface_above_ring และ stalk_surface_below_ring), ประเภทเยื่อหุ้มดอกเห็ดที่พบ (veil_type), สีของเยื่อที่หุ้มดอกเห็ด (veil_color), จำนวนวงแหวนของเห็ด (ring_number), รูปร่างของวงแหวน (ring_type), รอยช้ำบนเห็ด (bruise), กลิ่นของเห็ด (odor), สีรอยพิมพ์สปอร์ (spore_print_color), การเกาะกลุ่มกันของเห็ด (population), บริเวณที่พบเจอเห็ดเติบโต (habitat) และประเภทของเห็ด (class_ep) ดังตารางที่ 1

Table 1 Mushroom dataset

Variables (Abbreviation)	Possibilities
cap_color (CCO)	brown, buff, cinnamon, gray, green, pink, purple, red, white, yellow
cap_surface (CSU)	fibrous, grooves, scaly, smooth
cap_shape (CSH)	bell, conical, convex, flat, knobbed, sunken
gill_color (GCO)	black, brown, buff, chocolate, gray, green, orange, pink, purple, red, white, yellow
gill_attachment (GAT)	attached, descending, free, notched
gill_spacing (GSP)	close, crowded, distant
gill_size (GSI)	broad, narrow
stalk_shape (SSH)	enlarging, tapering
stalk_root (SRO)	bulbous, club, cup, equal, rhizomorphs, rooted, missing
stalk_color_above_ring (SCA)	brown, buff, cinnamon, gray, orange, pink, red, white, yellow
stalk_color_below_ring (SCB)	
stalk_surface_above_ring (SSA)	fibrous, scaly, silky, smooth
stalk_surface_below_ring (SSB)	
veil_type (VTY)	partial, universal
veil_color (VCO)	brown, orange, white, yellow
ring_number (RNU)	none, one, two
ring_type (RTY)	cobwebby, evanescent, flaring, large, none, pendant, sheathing, zone
bruise (BRU)	bruises, no
odor (ODO)	almond, anise, creosote, fishy, foul, musty, none, pungent, spicy
spore_print_color (SPC)	black, brown, buff, chocolate, green, orange, purple, white, yellow
population (POP)	abundant, clustered, numerous, scattered, several, solitary
habitat (HAB)	grasses, leaves, meadows, paths, urban, waste, woods
class_ep (CEP)	edible, poisonous

จากการสำรวจข้อมูลพบว่าตัวแปร “VTY” (ประเภทของเยื่อที่หุ้มดอกเห็ด) มีเพียงค่าเดียวคือ “partial” และพบว่าตัวแปร “SRO” (ลักษณะรากที่อยู่ปลายก้านดอก) มีค่าว่างมากกว่า 30% ของทั้งหมด ดังนั้นจึงลบ 2 ตัวแปรนี้ทิ้งไป และแทนค่าภายในตัวแปร “RNU” (จำนวนวงแหวนของเห็ด) ให้เป็นตัวเลขโดยตรง

จากนั้นกำหนดให้ “CEP” (ประเภทของเห็ด) เป็นลาเบล และอีก 20 ตัวแปรที่เหลือเป็นฟีเจอร์ที่ใช้ในการสร้างแบบจำลอง พร้อมกับการแปลงค่าให้เป็นตัวเลข โดยสำหรับฟีเจอร์ใช้เทคนิค Ordinal encoder ส่วนลาเบลใช้เทคนิค Label encoder แล้วจึงตรวจสอบค่าความสัมพันธ์ (correlation) ระหว่างข้อมูล

จากนั้นแบ่งชุดข้อมูลเป็นชุดข้อมูลสำหรับสร้างแบบจำลอง (Training set) และชุดข้อมูลสำหรับการทดสอบ (Test set) ทั้งหมด 3 อัตราส่วน คือ 50:50 (Training size = 50%), 60:40 (Training size = 60%) และ 70:30 (Training size = 70%) โดยกำหนดให้ในแต่ละอัตราส่วนมีเห็ดทั้งสองประเภทในจำนวนเท่าๆกัน แล้วปรับช่วงขอบเขตของฟีเจอร์ ด้วย “MinMaxscaler” ก่อนนำไปสร้างแบบจำลอง

2.2 การสร้างแบบจำลองเทคนิคการเรียนรู้ของเครื่องและปรับจูนหาค่า hyperparameters

งานวิจัยนี้ศึกษาทั้งหมด 5 แบบจำลอง โดยปรับจูน hyperparameters ด้วย RandomizedSearchCV โดยแต่ละแบบจำลองมีหลักการดังนี้

2.2.1 การถดถอยลอจิสติก (Logistic regression) หาค่าความน่าจะเป็น โดยใช้สมการ Sigmoid function (สมการที่ 1)

$$\sigma(x) = \frac{1}{1+e^{-x}} \quad (1)$$

2.2.2 ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine; SVM) สร้างเส้นแบ่งระหว่าง class (เส้น hyperplane) ของข้อมูลบน feature space เพื่อแบ่งข้อมูลออกเป็นกลุ่ม โดยเป็นเส้นที่มีขอบเขตห่างจาก support vector ที่เหมาะสม [8]

2.2.3 ต้นไม้ตัดสินใจ (Decision tree) มีการสร้างลำดับชั้นของการตัดสินใจขึ้นมาจากข้อมูลตัวแปรต้น โดยเริ่มจากเงื่อนไขแรกของการตัดสินใจ (Root node) แล้วแตกกิ่งการตัดสินใจไปที่เงื่อนไขถัดไป (Decision node) จนได้ผลลัพธ์สุดท้ายออกมา (Leaf node) ซึ่งแบบจำลองนี้จะแตกกิ่งจนได้ผลลัพธ์ที่ถูกต้องที่สุด [9]

2.2.4 แรนดอมฟอเรสท์ (Random Forest) เป็น Ensemble model ของ Decision Tree คือนำ Decision Tree หลายอัน มาทำนายผล ซึ่งใช้เทคนิค Bagging (Bootstrap aggregating) โดยแต่ละ Decision Tree นั้นมีการสุ่มฟีเจอร์และตัวอย่างที่ไม่เหมือนกัน ผลลัพธ์สุดท้ายของการทำนายนั้น ได้จากผลโหวตที่มากที่สุดของ Decision Tree แต่ละต้น [10]

2.2.5 เอ็กซ์ตรีมเกรเดียนต์บูสติ้ง (Extreme Gradient Boosting; XGBoost) เป็น Ensemble model ของ Decision Tree ด้วยเทคนิค Boosting คือนำ Decision Tree หลายอันมาเชื่อมกันเป็นลำดับ โดยแต่ละ Decision tree เรียนรู้จาก Error ของ Decision Tree ก่อนหน้า และจะหยุดเรียนรู้เมื่อไม่เหลือรูปแบบของ Error จาก Decision Tree ก่อนหน้าให้เรียนรู้อีก [11]

2.3 การคัดเลือกฟีเจอร์ (Feature selection)

งานวิจัยนี้ใช้ 2 เทคนิค คือ Recursive Feature Elimination (RFE) และ Chi-square โดยได้นำ hyperparameters ที่ได้จากขั้นตอนที่ 2.2 มาใช้ โดย RFE เป็นวิธีคัดเลือกฟีเจอร์โดยคำนวณจากการเรียนรู้และทดสอบ ในการสร้างแบบจำลองที่เลือกไว้ และ Chi-square เป็นวิธีคัดเลือกฟีเจอร์ก่อนนำเข้าสู่แบบจำลองเพื่อสร้างการเรียนรู้ โดยวิธีนี้เหมาะสมกับข้อมูลที่มีฟีเจอร์และลาเบลที่มีลักษณะเป็น category ซึ่งหากคะแนน chi-square มากแปลว่ามีความสำคัญมาก [12]

โดยทั้งสองเทคนิค ได้ศึกษาประสิทธิภาพกับชุดข้อมูลสำหรับทดสอบ เมื่อใช้จำนวนฟีเจอร์เท่ากับ 3, 5, 7 และ 9 ฟีเจอร์

2.4 การสกัดฟีเจอร์ (Feature extraction)

งานวิจัยนี้ใช้เทคนิคการวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis: PCA) ซึ่งเป็นการแปลงฟีเจอร์เดิมให้เป็นฟีเจอร์ใหม่ที่มีมิติลดลง [13] และศึกษาประสิทธิภาพกับชุดข้อมูลสำหรับทดสอบ เมื่อใช้จำนวนคอมโพแนนท์เท่ากับ 3, 5, 7 และ 9 คอมโพแนนท์ พร้อมกับการหาองค์ประกอบของฟีเจอร์ในแต่ละคอมโพแนนท์ (component loading)

2.5 การประเมินประสิทธิภาพของแบบจำลอง

งานวิจัยนี้ประเมินประสิทธิภาพของแบบจำลองเป็นค่า F1 score โดยหากค่า F1 score มีค่าสูงเข้าใกล้ 1 แปลว่าแบบจำลองมีประสิทธิภาพดี [14]

2.6 การพัฒนาต้นแบบเว็บแอปพลิเคชัน

งานวิจัยนี้ได้พิจารณาแบบจำลองและฟีเจอร์จากผลลัพธ์ที่ได้จากขั้นตอนที่ 2.3 เป็นหลัก โดยพิจารณาจากปัจจัยต่างๆ ได้แก่ ประสิทธิภาพ, เทคนิคและจำนวนฟีเจอร์ที่ใช้เหมาะสม โดยผู้วิจัยสร้างต้นแบบหน้าเว็บแอปพลิเคชัน ด้วยเครื่องมือ “Streamlit” ซึ่งเหมาะกับการสร้างหน้าเว็บแอปพลิเคชันสำหรับเทคนิคการเรียนรู้ของเครื่อง [15]

3. ผลการวิจัยและอภิปรายผลการวิจัย

3.1 ค่าความสัมพันธ์ (correlation) ระหว่างข้อมูล

พบว่า “GSI” (ขนาดของครีบเห็ด) มีความสัมพันธ์มากที่สุดกับ “CEP” (ประเภทของเห็ด) โดยมีค่าเท่ากับ 0.54 และฟีเจอร์ที่มีความสัมพันธ์กับประเภทของเห็ดรองลงมาคือ “SPC” (สีรอยพิมพ์สปอร์) และ “BRU” (รอยช้ำบนเห็ด) ดังรูปที่ 2

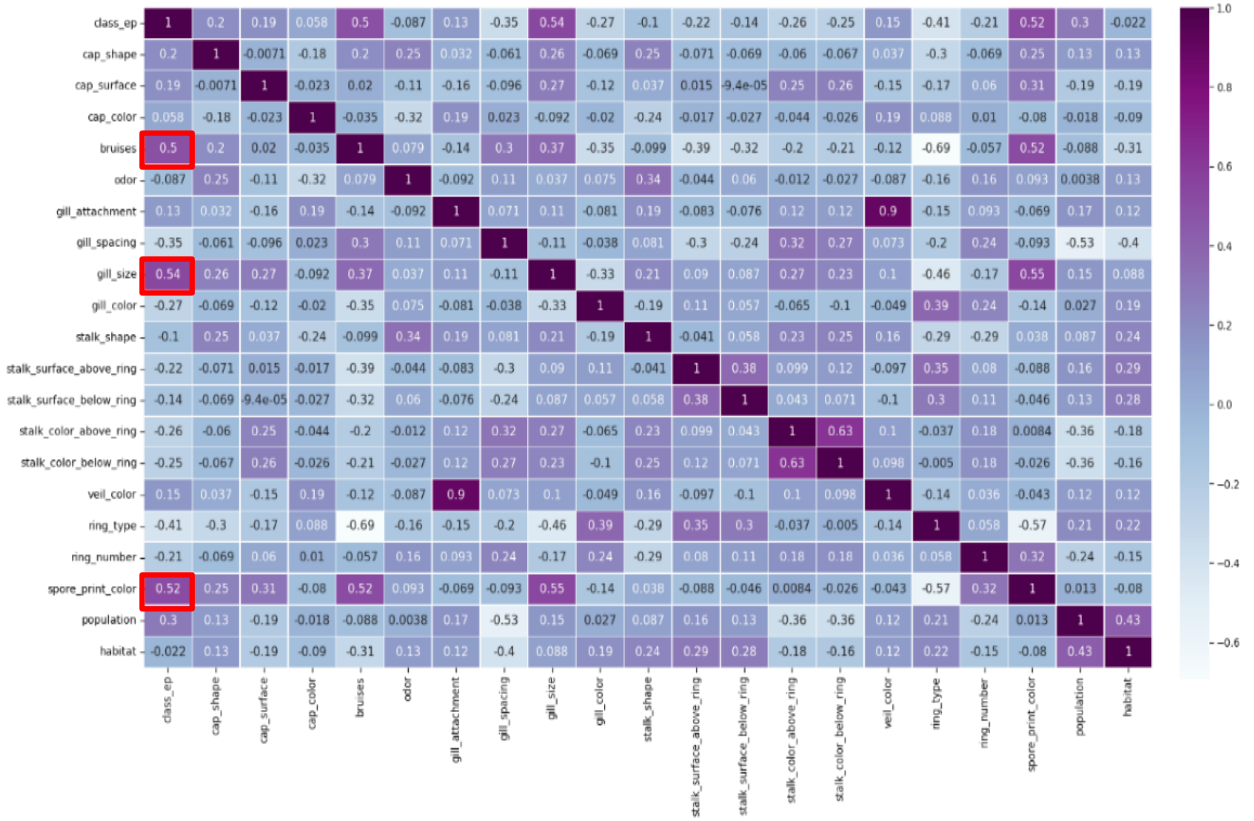

Figure 2 Correlation between features of mushroom data in ordinal type

3.2 ผลลัพธ์ของการสร้างแบบจำลองโดยใช้ 20 ฟีเจอร์

การสร้างแบบจำลองจำแนกประเภทเห็ดโดยใช้ฟีเจอร์ทั้ง 20 ฟีเจอร์พบว่าแบบจำลอง Logistic regression ให้ประสิทธิภาพที่ต่ำที่สุด เนื่องจาก Logistic regression เป็น linear model ซึ่งมีความสามารถในการประมวลผลข้อมูลที่มีความซับซ้อนมาก ได้ไม่ดีเมื่อเทียบกับแบบจำลองอื่นในงานวิจัยนี้ ส่วนแบบจำลอง Decision Tree ให้ค่า F1 score เท่ากับ 100% ยกเว้นที่ใช้ Training size = 60% ที่มีค่า F1 score เท่ากับ 99.69% เนื่องจาก hyperparameters ที่แบบจำลอง Decision Tree นั้นปรับค่าได้นั้นมีความแตกต่างกันในแต่ละ Training size ดังตารางที่ 2 และอาจเกิดจากความแตกต่างของการแบ่งชุดข้อมูลที่เกิดขึ้นในแต่ละ training size ทำให้แบบจำลอง Decision tree ที่ใช้ training size ที่แตกต่างกันเรียนรู้ได้ต่างกัน ส่วนอีก 3 แบบจำลอง คือ SVM, Random Forest และ XGBoost ให้ค่า F1 score เท่ากับ 100% ดังรูปที่ 3

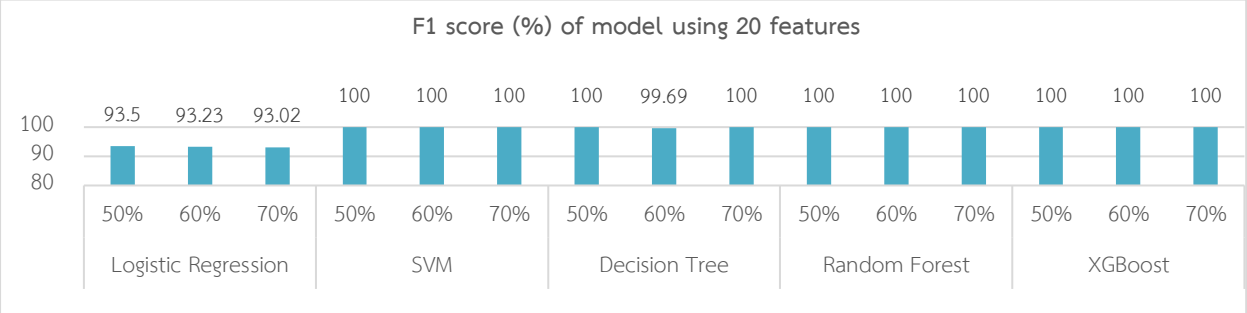

Figure 3 Performance metric (F1 score) of model using 20 features

Table 2 Parameters of Each model

Model	Training size	Parameters
Logistic Regression	All	solver = 'newton-cg', penalty = 'none', C = 22.456734693877554
SVM	All	kernel= 'poly', gamma = 5.0, degree=3.6666666666666665, C=73.68684210526315
Decision tree	50%	min_samples_split = 15, max_leaf_nodes = 46, max_features = 'sqrt', max_depth = 30
	60%, 70%	min_samples_split = 9, max_leaf_nodes = 35, max_features = 'log2', max_depth = 21
Random Forest	50%	n_estimators= 300, min_samples_split = 71, max_leaf_nodes = 57, max_features = 'log2', max_depth =52
	60%, 70%	n_estimators= 300, min_samples_split = 52, max_leaf_nodes = 43, max_features = 'log2', max_depth =85
XGBoost	50%, 60%	n_estimators=300, subsample = 0.5, min_child_weight= 1, max_depth= 3, learning_rate= 0.22749999999999998, colsample_bytree= 1
	70%	n_estimators= 300, subsample = 1, min_child_weight= 10, max_depth= 4, learning_rate= 0.3, colsample_bytree= 1

3.3 ผลลัพธ์จากการคัดเลือกฟีเจอร์

ผลโดยรวมของทั้งสองเทคนิคพบว่าแบบจำลอง Logistic Regression มีประสิทธิภาพน้อยกว่าแบบจำลองอื่น และสำหรับเทคนิค RFE พบว่า แบบจำลองส่วนใหญ่มีประสิทธิภาพที่ดี ส่วนแบบจำลอง XGBoost ที่ training size เท่ากับ 70% และ ใช้จำนวนฟีเจอร์เท่ากับ 3 ให้ประสิทธิภาพที่ต่ำที่สุดในการศึกษาทั้งหมด ซึ่งพบว่าฟีเจอร์ที่ได้นั้นค่อนข้างแตกต่างจากแบบจำลองอื่นที่ใช้จำนวนฟีเจอร์เท่ากัน (เท่ากับ 3) คือ ไม่พบว่าเลือกฟีเจอร์ชื่อ กลิ่นของเห็ด (ODO) หรือ สีรอยพิมพ์สปอร์ (SPC) มาใช้ในการประมวลผล โดยสามารถแสดงข้อมูลฟีเจอร์ได้ในตารางที่ 4 แต่เทคนิค RFE นี้มีข้อจำกัดกับแบบจำลองที่ไม่มีฟังก์ชั่นการหา Feature importance หรือค่า coefficient อย่าง SVM ที่ kernel เป็น “polynomial” นอกจากนี้พบว่าแบบจำลอง Decision tree และ Random Forest มีประสิทธิภาพใกล้เคียงกัน โดยเฉพาะที่จำนวนฟีเจอร์เท่ากับ 3 พบว่ามีประสิทธิภาพเท่ากันในทุก training size

เมื่อเปรียบเทียบผลลัพธ์ระหว่างเทคนิค Chi-square กับเทคนิค RFE นั้นพบว่าที่จำนวนฟีเจอร์น้อยๆ การใช้เทคนิค RFE มีประสิทธิภาพดีกว่า เนื่องจาก RFE เป็นเทคนิคที่คำนวณจากการเรียนรู้ในการสร้างแบบจำลอง ในขณะที่ Chi-square เป็นเทคนิคที่คัดเลือกฟีเจอร์ก่อนเข้าสู่แบบจำลอง ซึ่งจากกราฟในรูปที่ 4 เห็นได้ว่าที่ฟีเจอร์เท่ากับ 3 ของทุกๆ training size นั้น การใช้เทคนิค RFE ทำให้ F1 score ของแบบจำลองส่วนใหญ่มีค่ามากกว่า 99% ในขณะที่การใช้เทคนิค chi-square นั้น แบบจำลองต่างๆ มีค่า F1 score น้อยกว่า 90% เนื่องจากฟีเจอร์ที่ถูกคัดเลือกในแต่ละเทคนิคนั้นมีความแตกต่างกัน ดังแสดงในตารางที่ 3-4

3.4 ผลลัพธ์จากการสกัดฟีเจอร์

แบบจำลอง SVM ที่ใช้จำนวนคอมโพแนนท์เท่ากับ 9 ให้ค่า F1 score ที่มากที่สุด คือ 100% เนื่องจาก SVM ที่มี kernel เป็น Polynomial สามารถประมวลผลในการใช้ฟีเจอร์ที่มีมิติสูงๆ ได้ดีและพบว่า Logistic Regression ให้ประสิทธิภาพที่ค่อนข้างน้อยกว่าแบบจำลองอื่น ดังรูปที่ 4 และสามารถแสดงองค์ประกอบของฟีเจอร์ในแต่ละคอมโพแนนท์ (component loading) ได้ตามรูปที่ 5 โดยในแต่ละองค์ประกอบของฟีเจอร์ในแต่ละคอมโพแนนท์ ฟีเจอร์ที่มีค่าเป็นบวก (สีเหลือง) แปลว่าเป็นฟีเจอร์ที่มีความสำคัญมากโดยให้ผลเชิงบวกต่อคอมโพแนนท์นั้นๆ มาก และฟีเจอร์ที่ให้ค่าเป็นลบ (สีน้ำเงินเข้ม) แปลว่าเป็นฟีเจอร์ที่มีความสำคัญมากโดยให้ผลเชิงลบต่อคอมโพแนนท์นั้นๆ มาก เช่น ฟีเจอร์ “BRU” นั้นให้ผลเชิงบวกในคอมโพแนนท์ที่ 1 ของทุกการใช้ n_components ส่วนฟีเจอร์ “SSH” ให้ผลเชิงลบในคอมโพแนนท์ที่ 1 ของทุกการใช้ n_components นอกจากนี้ในคอมโพแนนท์อื่นๆ ยังพบฟีเจอร์ที่ให้ผลเชิงบวกได้แก่ “CSU”, “GCO”, “GSP”, “SSH” เป็นต้น และพบฟีเจอร์ที่ให้ผลเชิงลบได้แก่ “CCO”, “SSH”, “CSU”, “SSB” เป็นต้น

Table 3 Feature names of feature selection (Chi-square)

No. of features	Model	Training size	Feature names (as Abbreviation)
3	All	All	GSI, BRU, GSP
5	All	All	GSI, BRU, GSP, SPC, RTY
7	All	All	GSI, BRU, GSP, SPC, RTY, GCO, CSU
9	All	All	GSI, BRU, GSP, SPC, RTY, GCO, CSU, POP, SCA

Table 4 Feature names of feature selection (RFE)

No. of features	Model	Training size	Feature names (as Abbreviation)
3	Logistic regression	All	VCO, RNU, SPC
	XGBoost	50%, 60%	
		70%	SSH, VCO, RNU
	Random Forest	All	ODO, GSI, SPC
	Decision tree	50%, 60%	
		70%	ODO, SPC, POP
5	Logistic regression	All	BRU, GSP, VCO, RNU, SPC
	XGBoost	50%, 60%	ODO, GSI, VCO, RNU,SPC
		70%	GSP, SSH, VCO, RNU, SPC
	Random Forest	All	ODO, GSI, RTY, SPC, POP
	Decision tree	50%	ODO, SPC, SCB, POP, HAB
		60%	CCO, ODO, GSI, SPC, POP
		70%	ODO, GSP, GSI, SCA, HAB
7	Logistic regression	All	BRU, GSP, SSA, VCO, RTY, RNU, SPC
	XGBoost	50%	ODO, GSP, GSI, VCO, RNU, SPC, POP
		60%	ODO, GSP, GSI, SSA, VCO, RNU, SPC
		70%	ODO, GSP, GSI, SSH, VCO, RNU, SPC
	Random Forest	All	BRU, ODO, GSP, GSI, RTY, SPC, POP
	Decision tree	50%	BRU, ODO, GSI, SSA, SSB, RTY, SPC
		60%	BRU, ODO, GSI, SSB, SCB, SPC, POP
		70%	ODO, GSP, GSI, SCA, RTY, SPC, POP
9	Logistic regression	All	CSU, BRU, GSP, GSI, SSA, VCO, RTY, RNU, SPC
	XGBoost	50%, 60%	ODO, GSP, POP, SSA, VCO, SSB, GSI, RNU, SPC
		70%	CSU, ODO, GSI, GSP, SSH, VCO, RNU, SPC, POP
	Random Forest	50%, 60%	BRU, ODO, GSP, GSI, RTY, SSA, SSB, SPC, POP
		70%	BRU, ODO, GSP, GSI, RTY, SSB, HAB, SPC, POP
	Decision tree	50%	BRU, GSP, GSI, SSH, HAB, SSB, SCB, SPC, POP
		60%	ODO, SSH, HAB, SSB, SCA, RTY, RNU, SPC, POP
		70%	BRU, ODO, GSP, GSI, SSH, HAB, RTY, SPC, SCA

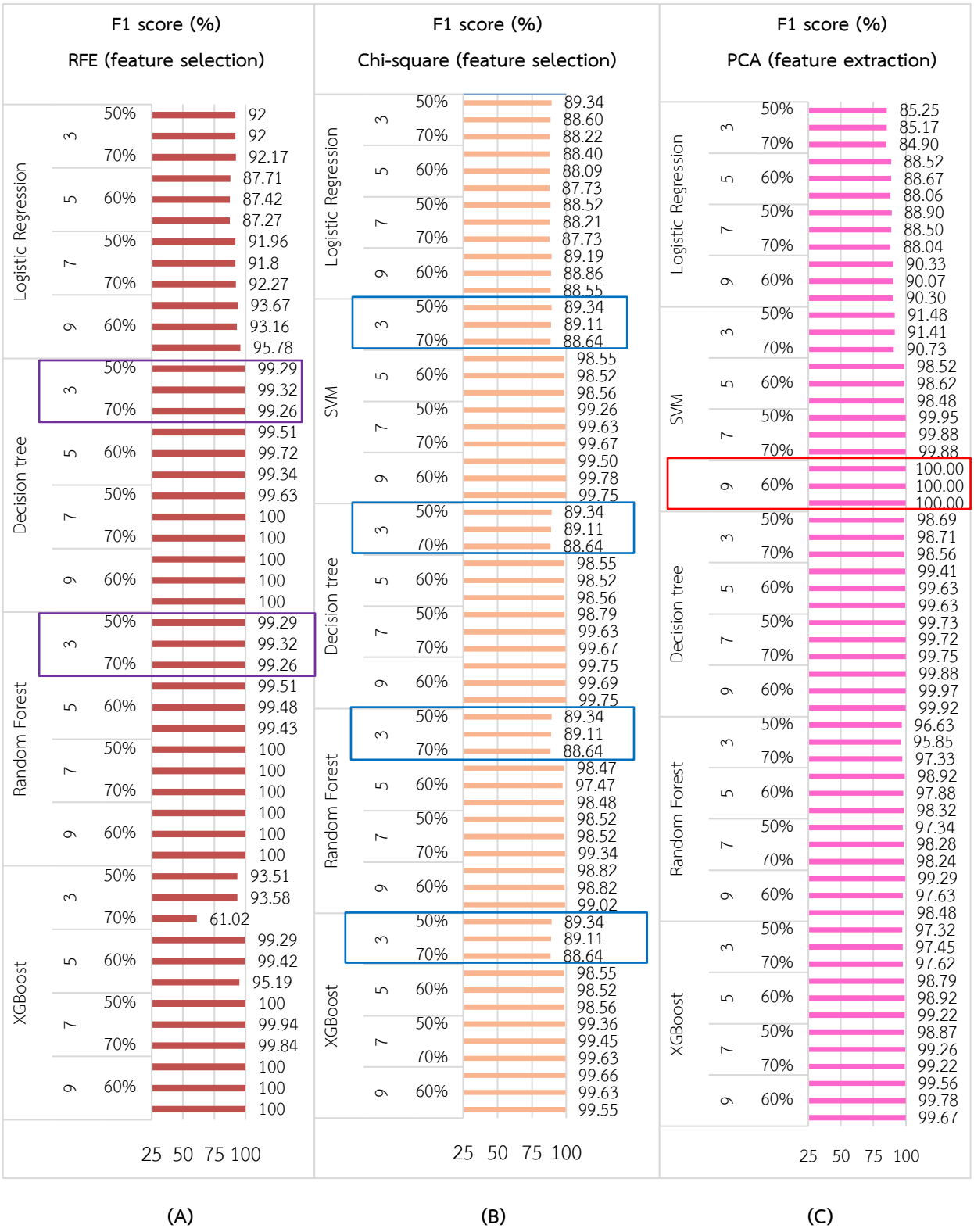

Figure 4 Performance metric of model: (A) Feature selection (RFE), (B) Feature selection (chi-square) and (C) Feature extraction (PCA)

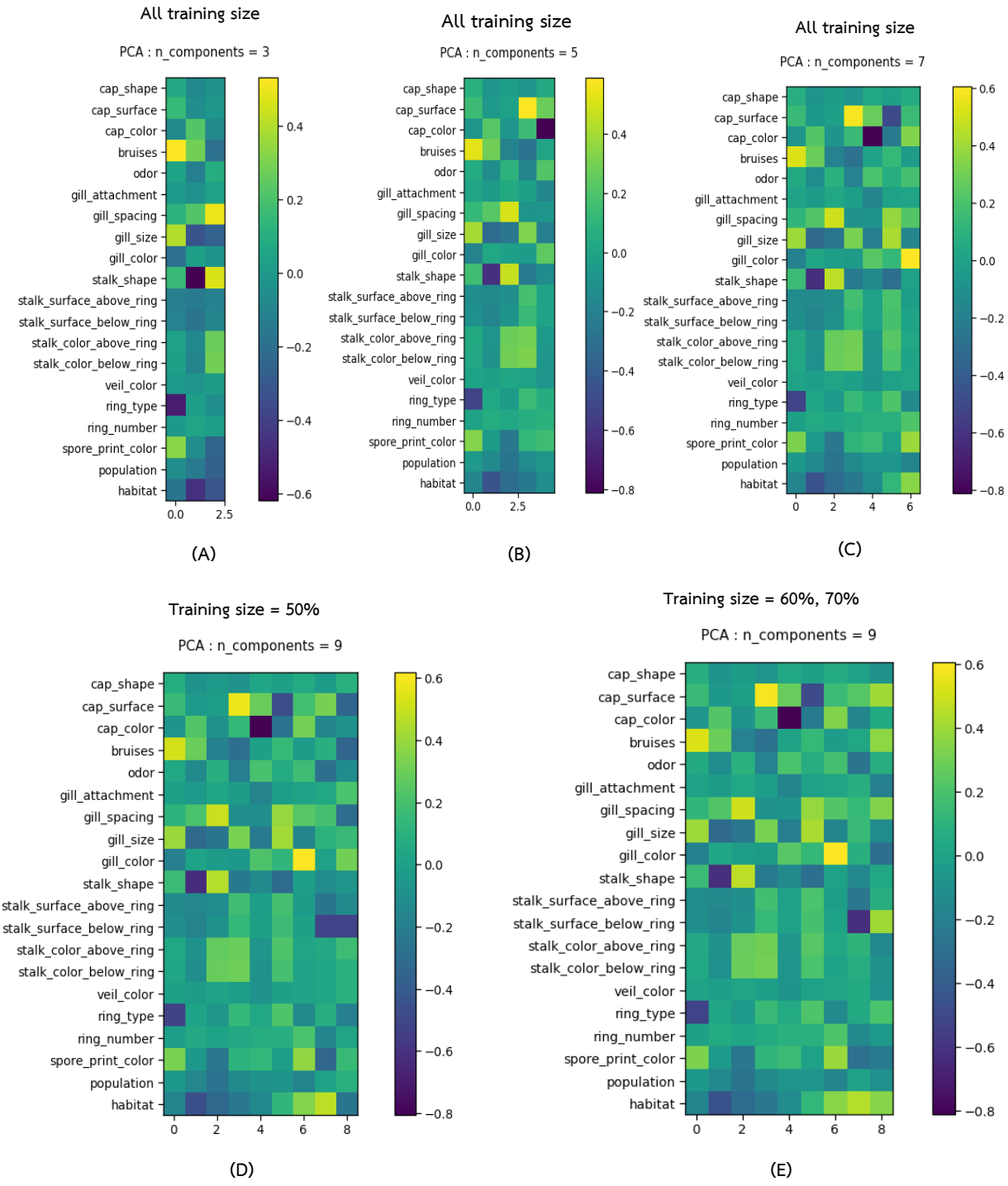

Figure 5 Component Loading of PCA technique in each n_components and training size: (A) n_components = 3 of all training size, (B) n_components = 5 of all training size, (C) n_components = 7 of all training size, (D) n_components = 9 of training size = 50% and (E) n_components = 9 of training size = 60%, 70%

3.5 การพัฒนาต้นแบบเว็บแอปพลิเคชัน

จากผลการศึกษาของการคัดเลือกฟีเจอร์ เมื่อพิจารณาจากทั้งประสิทธิภาพของแบบจำลอง, เทคนิคการคัดเลือก/การสกัดฟีเจอร์, training size ที่ใช้ และจำนวนฟีเจอร์ พบว่าแบบจำลอง Decision tree ที่ใช้เทคนิค RFE, training size เท่ากับ 60% และใช้จำนวนฟีเจอร์เท่ากับ 3 ฟีเจอร์ ซึ่งสามารถแสดงชื่อฟีเจอร์คือ กลิ่นของเห็ด (ODO), ขนาดครีบเห็ด (GSI) และสีรอยพิมพ์สปอร์ (SPC) มีความเหมาะสมที่สุด (F1 score = 99.32%) โดยสามารถแสดงต้นแบบเว็บแอปพลิเคชันได้ตามรูปที่ 6

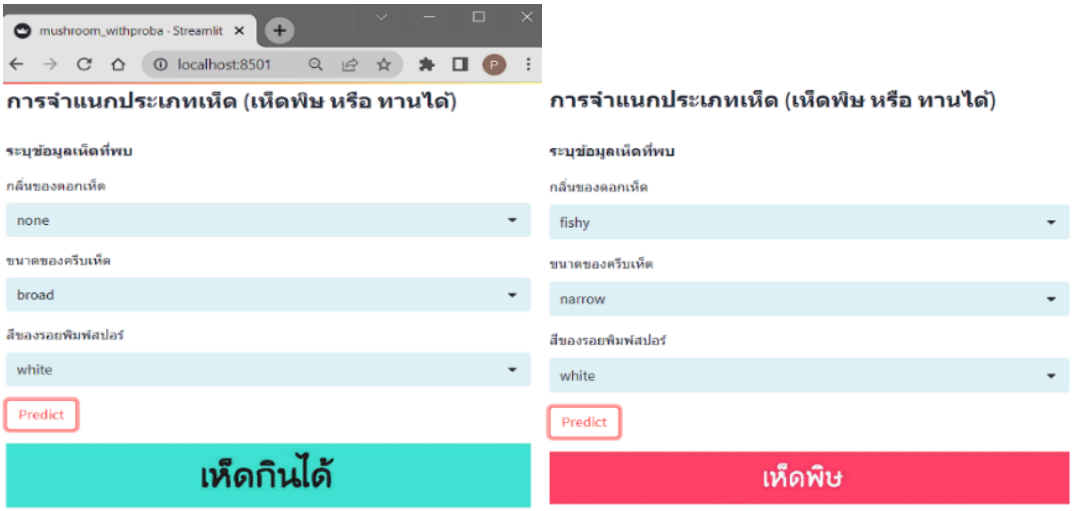

Figure 6 Example of prototype of web application

4. สรุป

จากผลการศึกษาการจำแนกประเภทของเห็ดระหว่างเห็ดพิษและเห็ดที่รับประทานได้ โดยใช้ชุดข้อมูลสาธารณะของเห็ดครีบจากหนังสือข้อมูลเห็ดในฝั่งอเมริกาเหนือ ปี 1981 [7] พบว่าการศึกษาที่ให้ประสิทธิภาพที่เหมาะสมที่สุด คือ แบบจำลอง Decision tree ร่วมกับการคัดเลือกฟีเจอร์ด้วยเทคนิค RFE ที่ใช้ training size เท่ากับ 60% และใช้ฟีเจอร์จำนวน 3 ฟีเจอร์คือ กลิ่นของเห็ด (ODO), ขนาดครีบเห็ด (GSI) และสีรอยพิมพ์สปอร์ (SPC) ซึ่งให้ค่า F1 score = 99.32% ซึ่งใกล้เคียงและสอดคล้องกับงานวิจัยของ Vanitha และคณะ [6] ที่มี F1 score = 99%

อย่างไรก็ตามการวิจัยนี้ใช้ข้อมูลของเห็ดจากแหล่งข้อมูลดั้งเดิมคือเป็นเห็ดที่พบในฝั่งอเมริกาเหนือ ซึ่งอาจแตกต่างกับเห็ดภายในประเทศไทย ดังนั้นการศึกษาครั้งถัดไปจึงควรมีการเก็บข้อมูลของเห็ดที่พบภายในประเทศไทยเพิ่มเติม และควรมีการศึกษาแบบจำลองอื่นเพิ่มเติม ได้แก่ k-nearest neighbors, Naïve Bayes, VotingClassifier หรือ AutoSklearnClassifier (แบบจำลองที่มีการปรับจูนหาแบบจำลองและ hyperparameters ที่เหมาะสมอย่างอัตโนมัติ) หรือนอกจากนี้อาจนำรูปภาพมาใช้ในการประมวลผลภาพ (Image processing) เพื่อใช้ในการจำแนกประเภทเห็ด ร่วมกับการใช้แบบจำลองโครงข่ายประสาทเทียม (Neural Network) หรือเทคนิคการเรียนรู้เชิงลึก (Deep learning techniques)

5. กิตติกรรมประกาศ

ขอขอบคุณคณาจารย์ทุกท่านในภาควิชาวิทยาการข้อมูล มหาวิทยาลัยศรีนครินทรวิโรฒ ที่แนะนำความรู้และแนวทางในการทำงานวิจัยเรื่องนี้ และขอขอบคุณบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒที่สนับสนุนการเผยแพร่การวิจัยนี้

6. References

[1] Sakonrak, B., Duangkae, K., Ayawong, J., Pongpanich, K. and Khemaman, W., 2006, Gilled Mushrooms: Kaeng Krachan Forest Complex Chiang Dao Wildlife Sanctuary and Phu Khieo Wildlife Sanctuary, Forest Conservation Research Division Forest and Plant Conservation Research Office Department of National Parks, Wildlife and Plant Conservation, Bangkok, 240 p. (in Thai)

[2] Taeprasert, N., Poisonous Mushrooms, 2009, Office of Disease Prevention and Control 5 Nakhon Ratchasima, Department of Disease Control, Ministry of Public Health, Nakhon Ratchasima, 67 p. (in Thai)

[3] Islam, M.D., 2015, Cultivation Techniques of Edible Mushrooms: Agaricus bisporus, Pleurotus spp., Lentinula edodes and Volvariella volvocea, Available source: https://www.researchgate.net/publication/275179411_Cultivation_Techniques_of_Edible_Mushrooms_Agaricus_bisporus_Pleurotus_spp_Lentinula_edodes_and_Volvariella_volvocea, Sep 22, 2022.

[4] Verma, S. and Dutta, M., 2018, Mushroom Classification Using ANN and ANFIS Algorithm, IOSR Journal of Engineering (IOSRJEN), 8(1): 26-32.

[5] Chitayae, N. and Sunyoto, A., 2020, Performance Comparison of Mushroom Types Classification Using K-Nearest Neighbor Method and Decision Tree Method, In 2020 3rd International Conference on Information and Communications Technology (ICOIACT), pp. 308-313. IEEE.

[6] Vanitha, V., Ahil, M.N. and Rajathi N., 2020, Classification of mushrooms to detect their edibility based on key attributes, Bioscience Biotechnology Research Communications, 13(11): 37-41.

[7] Schlimmer, J., Mushroom Data Set, Machine Learning Repository, Available source: https://archive.ics.uci.edu/ml/datasets/Mushroom, Sep 22, 2022.

[8] Agarwal, D., 2021, Introduction to SVM(Support Vector Machine) Along with Python Code, Data Science Blogathon, Available source: https://www.analyticsvidhya.com/blog/2021/04/insight-into-svm-support-vector-machine-along-with-code/, Oct 12, 2022.

[9] Saini, A., 2021, Decision Tree Algorithm – A Complete Guide, Data Science Blogathon, Available source: https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/, Oct 12, 2022

[10] What is a Random Forest?, TIBCO Software Inc, Available source: https://www.tibco.com/reference-center/what-is-a-random-forest, Oct 12, 2022.

[11] mariajesusbigml. Introduction to Boosted Trees, Available source: https://blog.bigml.com/2017/03/14/introduction-to-boosted-trees/, Oct 12, 2022.

[12] Brownlee, J., How to Choose a Feature Selection Method For Machine Learning, Machine Learning Mastery, Available source: https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/, Oct 12, 2022.

[13] Hira, Z.M. and Gillies, D.F., 2015, A Review of feature selection and feature extraction methods applied on microarray data, Advances in Bioinformatics, 2015: 1-13.

[14] Sagrawal, S.K., Metrics to Evaluate your Classification Model to take the right decisions, Analytics Vidhya, Available source: https://www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your-classification-model-to-take-the-right-decisions/, Oct 12, 2022.

[15] Streamlit documentation, Streamlit Inc, Available source: https://docs.streamlit.io/, Feb 15, 2023.