Term Project

2024. 12. 09

Khin Wai Yan, Suhan Jang, Yenah Cho

https://colab.research.google.com/drive/1dhxF-Uwwu7PMqqOTIp_7agwZAOxGkiF1?usp=sharing

Contents

Goal

Classify mushrooms into two catgories:

edible (class 0) and poisonous (class 1)

- Which machine learning models perform best on this dataset?
- Which features are most indicative of a poisonous mushroom?

Data

`mushroom.csv`

(UCI Machine Learning Repository)

Preprocessing

- Modal imputation
- one-hot encoding
- z-score normalization
- feature selection

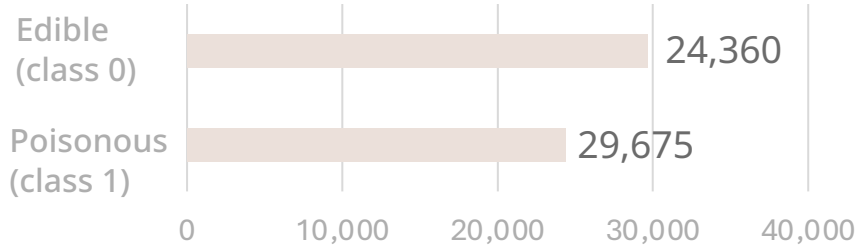
`mushroom_cleaned.csv`

9 columns

- Cap Diameter
- Cap Shape
- Gill Attachment
- Gill Color
- Stem Height
- Stem Width
- Stem Color
- Season
- Target Class
  - `0:` Edible
  - `1:` Poisonous

Data Stats

Total instances: 54,035



Models

01 Random Forest

02 LightGBM

Light Gradient Boosting Machine

03 XGBoost

eXtreme Gradient Boosting

04 MLP

Multi-Layer Perceptron

05 H2O – AutoML

GBM and DRF

06 Clustering

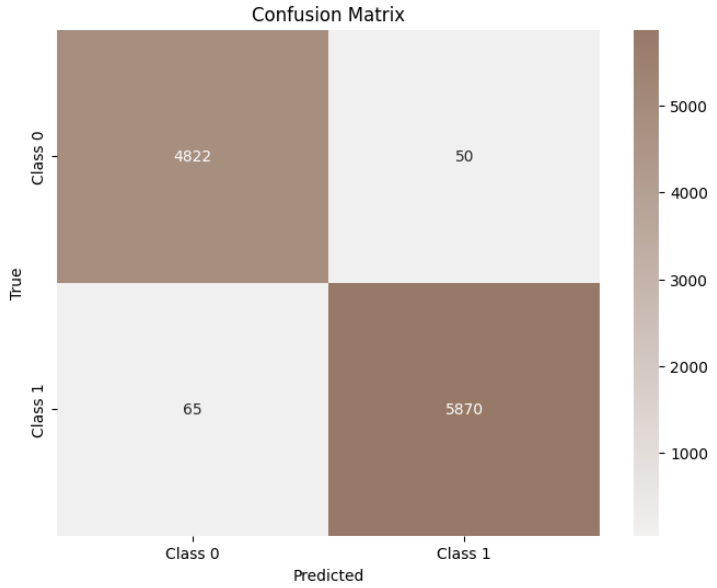
K-means and Agglomerative

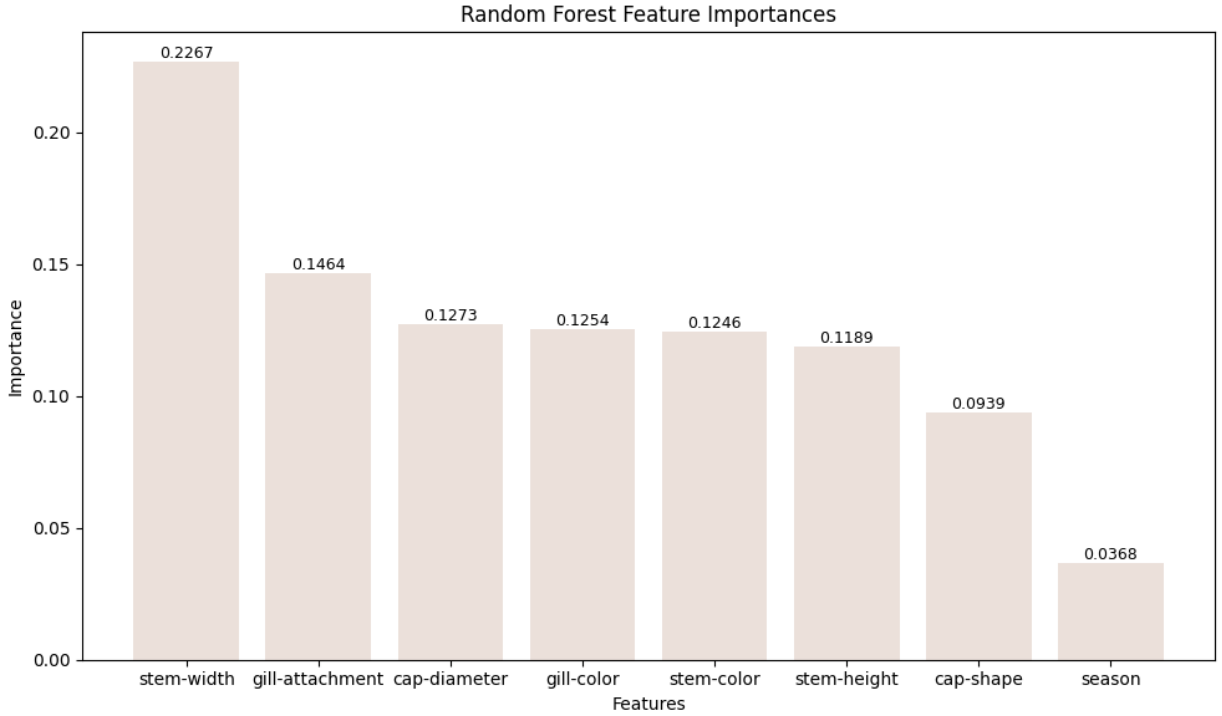
Results > Random Forest

Accuracy: 0.9893587489590081

Classification Report:

	precision	recall	f1-score	support
0	0.99	0.99	0.99	4872
1	0.99	0.99	0.99	5935
accuracy			0.99	10807
macro avg	0.99	0.99	0.99	10807
weighted avg	0.99	0.99	0.99	10807



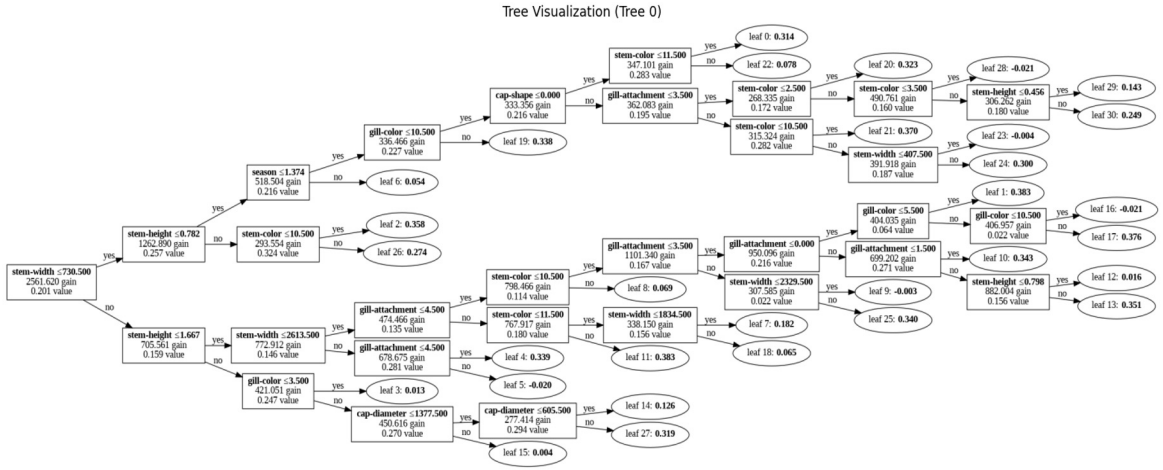


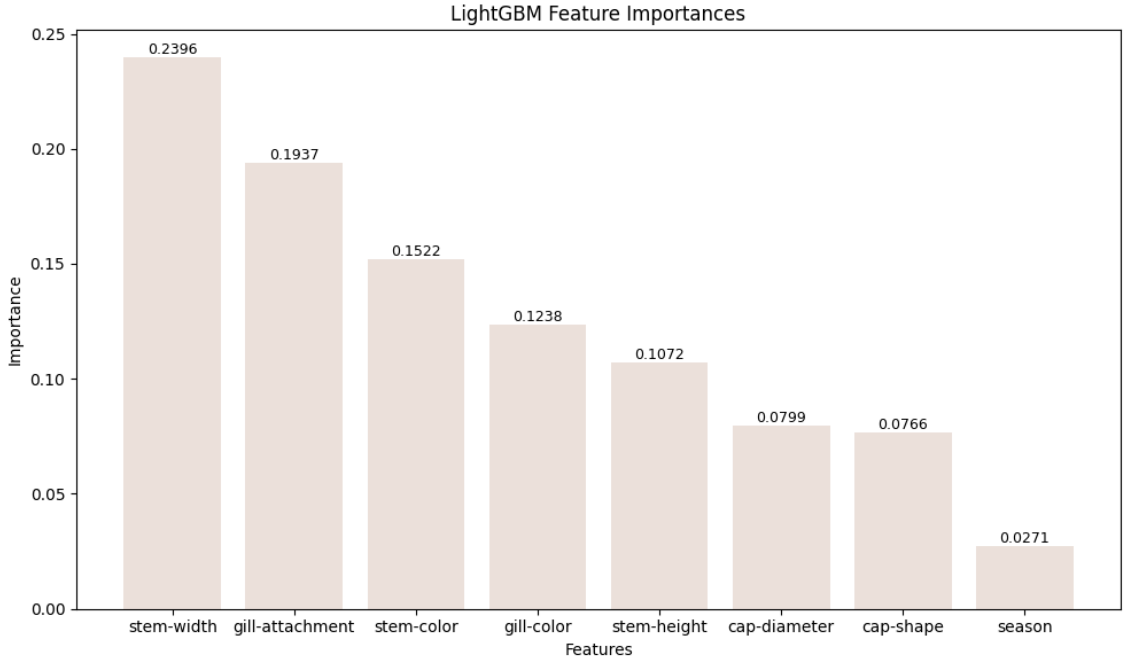
Results > LightGBM

Accuracy: 0.9773295086517998

Classification Report:

	precision	recall	f1-score	support
0	0.97	0.98	0.98	4909
1	0.98	0.98	0.98	5898
accuracy			0.98	10807
macro avg	0.98	0.98	0.98	10807
weighted avg	0.98	0.98	0.98	10807

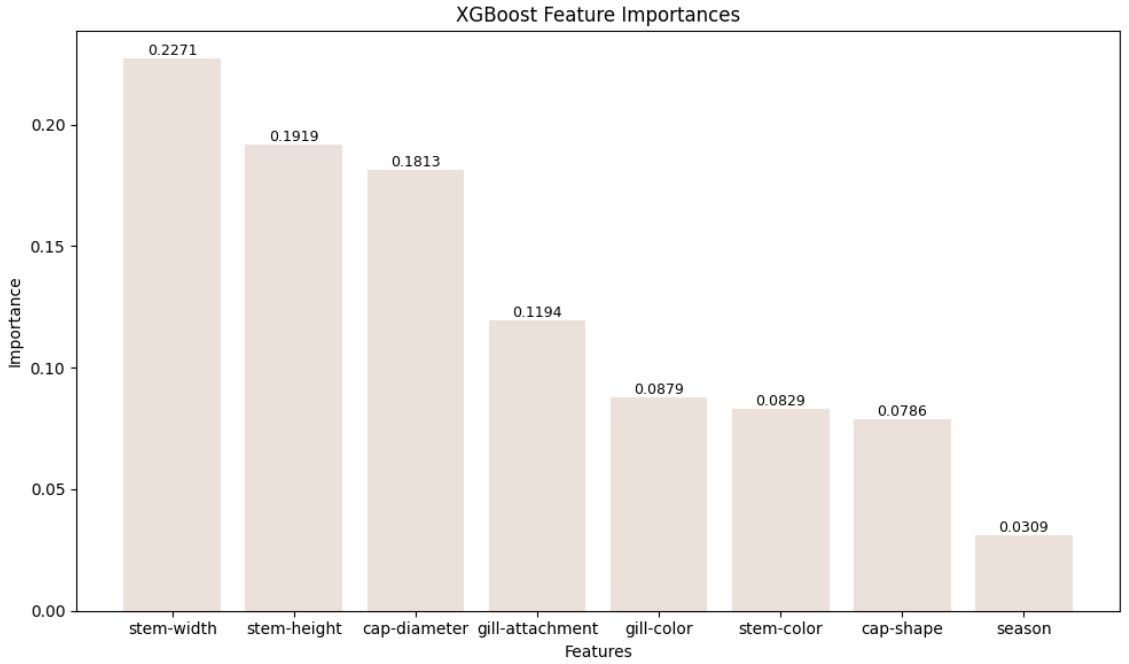




Results > XGBoost

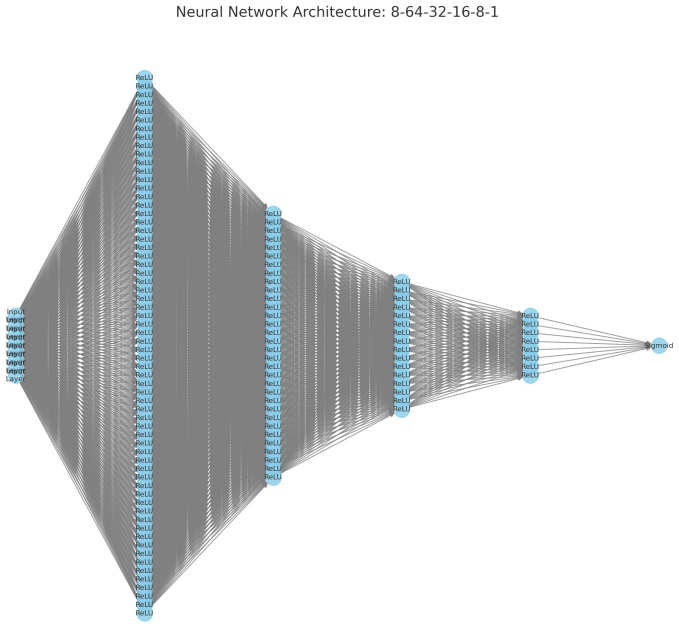
- max_depth = 5
- eta= = 0.1
- Binary: logistics, log loss
- num_round = 1500

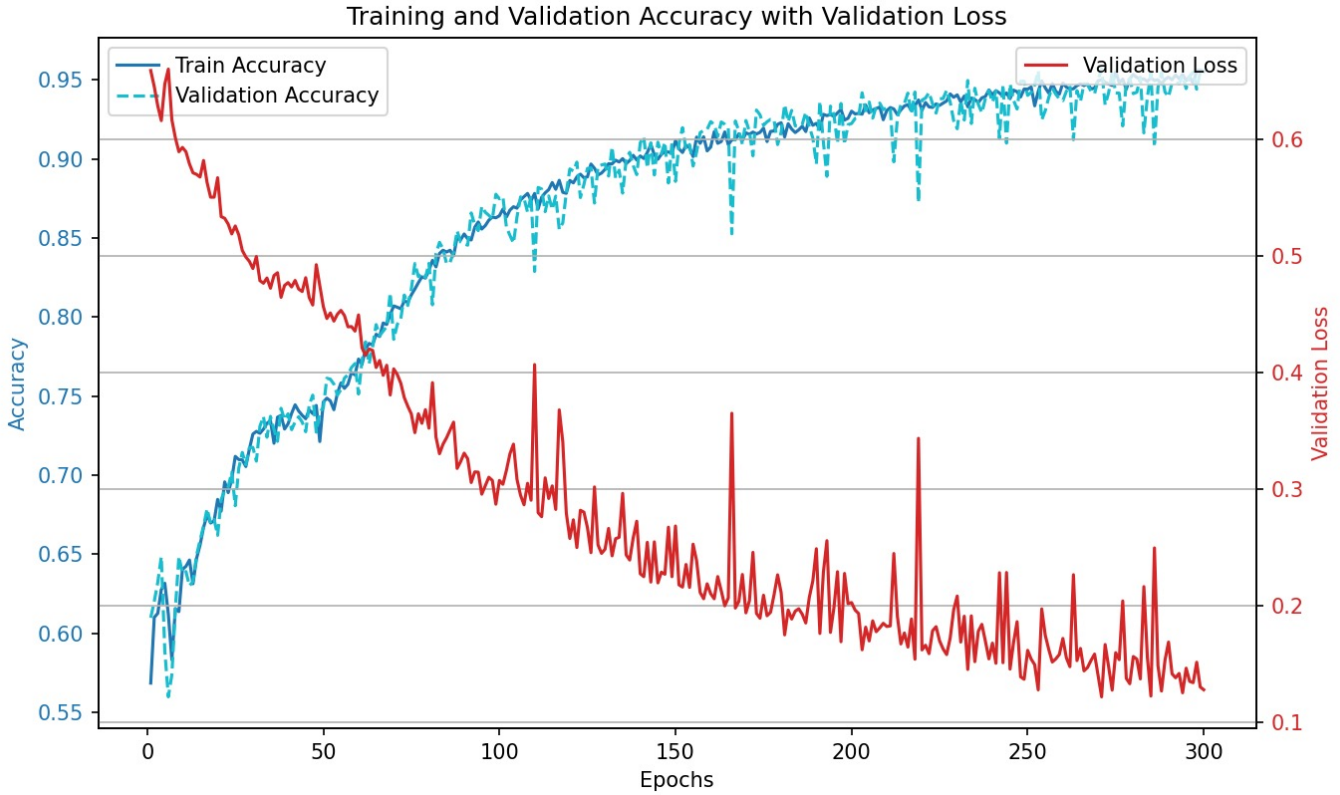
```
Accuracy: 0.9904691403719811

Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      4909
           1       0.99      0.99      0.99      5898

    accuracy                           0.99     10807
   macro avg       0.99      0.99      0.99     10807
weighted avg       0.99      0.99      0.99     10807
```



Results > MLP

- 200 epochs
- Batch size = 32
- 8 – 64 – 32 – 16 – 8 - 1





Test Accuracy: 0.9573424458503723

Results > H2O - AutoML

Top Model

Model type: gbm

- F1 score: 0.990895
- Accuracy: 0.989997
- ntrees: 234
- max_depth: 8
- min_rows: 10.0
- sample_rate: 0.8
- learn_rate: 0.1

Model type: drf

- F1 score: 0.991483
- Accuracy: 0.990646
- ntrees: 46
- max_depth: 20
- min_rows: 1.0
- sample_rate: 0.632

Discussion

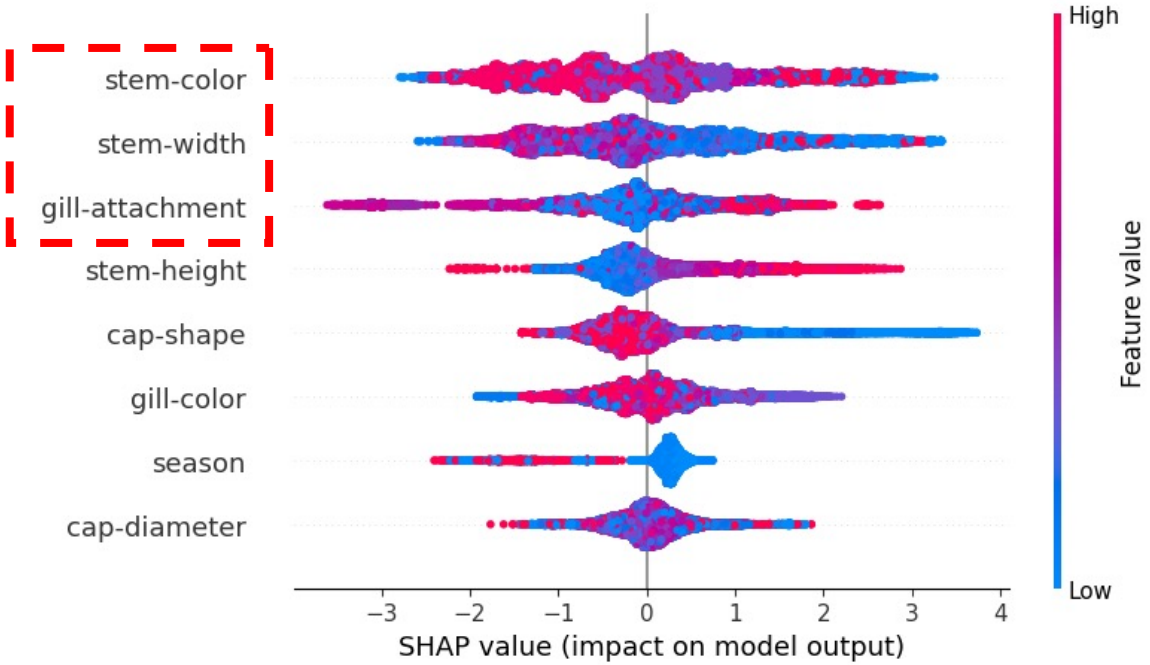
Which machine learning models perform best on this dataset?

Model	Accuracy (2 d.p)	Top 3 Feature Importances
H2O – AutoML (gbm)	**99.09%**	-
H2O – AutoML (drf)	**99.15%**	-
XGBoost	99.05%	**stem-width**, stem-height, cap-diameter
Random Forest	98.94%	**stem-width**, gill-attachment, cap-diameter
LightGBM	97.73%	**stem-width**, gill-attachment, stem-color
MLP	95.73%	-

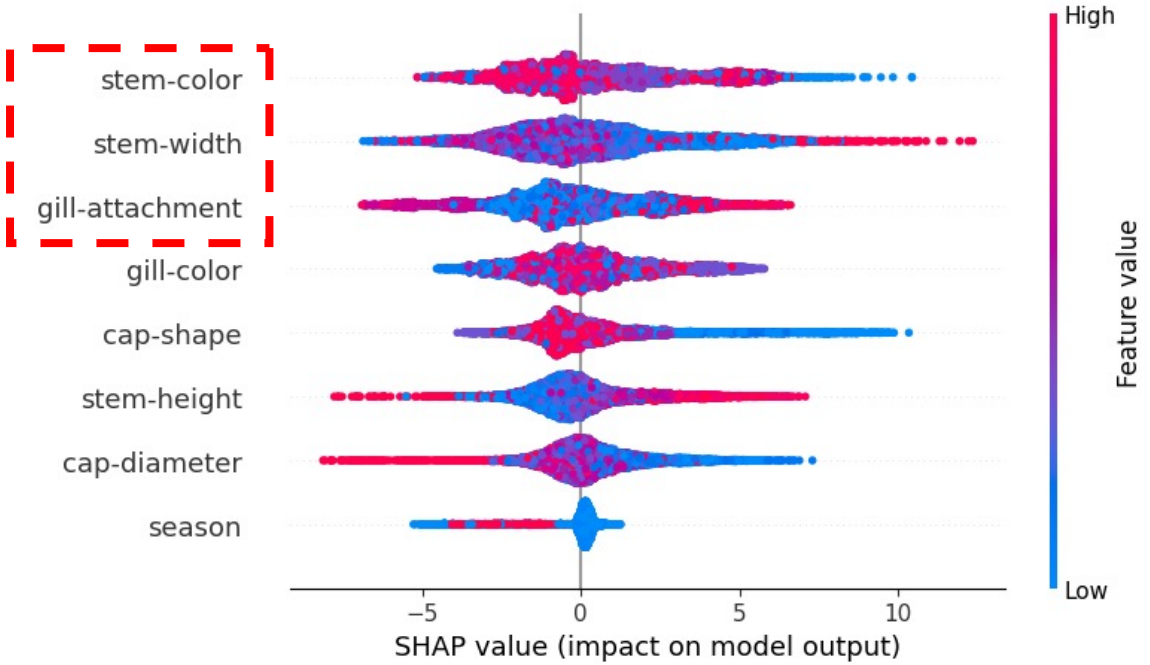
SHAP Analysis > Poisonous or edible?

Which features are most indicative of a poisonous mushroom?

LightGBM

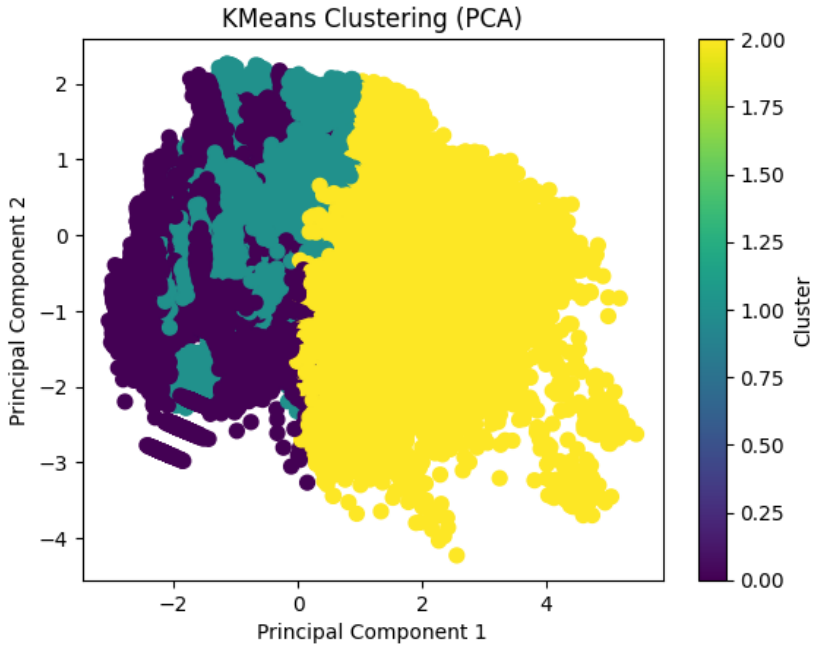


XGBoost



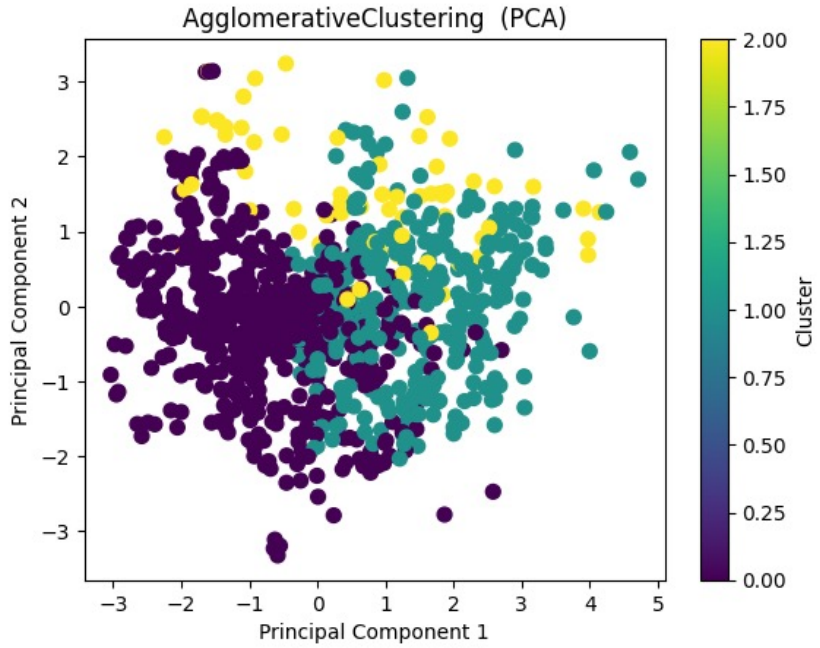
[Extension] Clustering

K-means clustering




```
Silhouette Score: 0.14324906421096753
Adjusted Rand Index (ARI): 0.014693770636364296
Normalized Mutual Information (NMI): 0.014536987048463349
```

Agglomerative clustering




```
Silhouette Score: 0.16697595841941235
Adjusted Rand Index (ARI): 0.04173505561048274
Normalized Mutual Information (NMI): 0.032487349049258776
```

Thank you!