Northwestern

Predicting Mushroom Edibility: Classification & Clustering Analysis

Data Mining Final Project - Winter 2025

Team 02: Haowen Geng, Runxuan Li,
Ruyi Lu, Wenfei Wu, Wenyang Miao

Executive Summary

Identifying Mushroom Edibility and Grouping Patterns Using Machine Learning Methods

Outcome: A data-driven approach to **classify mushrooms as edible or poisonous** with high accuracy, while **identifying natural groupings** based on ecological and physical characteristics.

Problems

Class Imbalance
Majority of mushrooms are poisonous

Data Complexity
Most features are categorical

Feature Importance
May strongly influence edibility

Cluster Interpretability
Need to have known ecological groups

Strategies

Machine Learning Strategies to Implement:

Feature Engineering
One-hot encoding

Data Handling
Oversampling (SMOTE)

Dimensionality Reduction
PCA & t-SNE

Clustering Methods
K-Means, DBSCAN, and Hierarchical Clustering

Analysis

Classification Performance: Random Forest & XGBoost models achieved high accuracy (~98%).

- **Key Features:** Spore print color, gill attachment, and cap surface texture.

Clustering Insights: K-Means clustering (k=4) provided meaningful groupings.

- Some clusters aligned with known ecological and taxonomic categories of mushrooms.

Dataset Overview & Data Cleaning Steps

Discuss project scope and preliminary steps taken

The mushroom dataset consists of 61,069 mushrooms with 21 features (categorical & numerical).

Two key objectives:

1. **Clustering:** Group mushrooms based on **physical and ecological traits**.
2. **Classification:** Predict whether a mushroom is **edible (e)** or **poisonous (p)**.

Steps Performed: Data Cleaning

1. Handling Missing Values

- Missing categorical values were imputed using the **mode** (most frequent category).
- Some missing values were treated as a **separate category** to retain information.

2. Encoding Categorical Variables

- **One-hot encoding** was applied.
- **Ordinal encoding** considered for ordered variables.

3. Addressing Class Imbalance

- **Oversampling: SMOTE** for synthetic samples

4. Feature Standardization

- **Numerical features** were **standardized** for uniform scaling.
  - **Min-Max Scaling**
  - **Z-score normalization**

5. Outlier Detection

- **Extreme values** were checked to ensure not data entry errors.

Exploratory Data Analysis

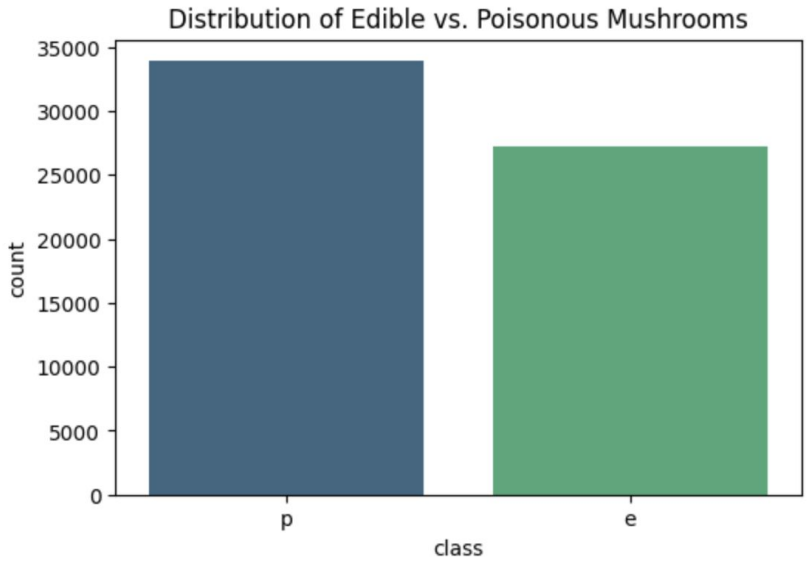
Missing value and data balance

The dataset has **21 features**, with significant missing values in some columns.

Features with the **most missing values**:

- **Veil-type (57,892 missing)**
- **Spore-print-color (54,715 missing)**
- **Stem-root (51,538 missing)**

```
Missing Values:
cap-diameter                0
cap-shape                   0
cap-surface             14120
cap-color                   0
does-bruise-or-bleed        0
gill-attachment          9884
gill-spacing            25063
gill-color                  0
stem-height                 0
stem-width                  0
stem-root               51538
stem-surface            38124
stem-color                  0
veil-type               57892
veil-color              53656
has-ring                    0
ring-type                2471
spore-print-color       54715
habitat                     0
season                      0
class                       0
dtype: int64
```



The dataset contains **more poisonous (p) mushrooms** than edible (e).

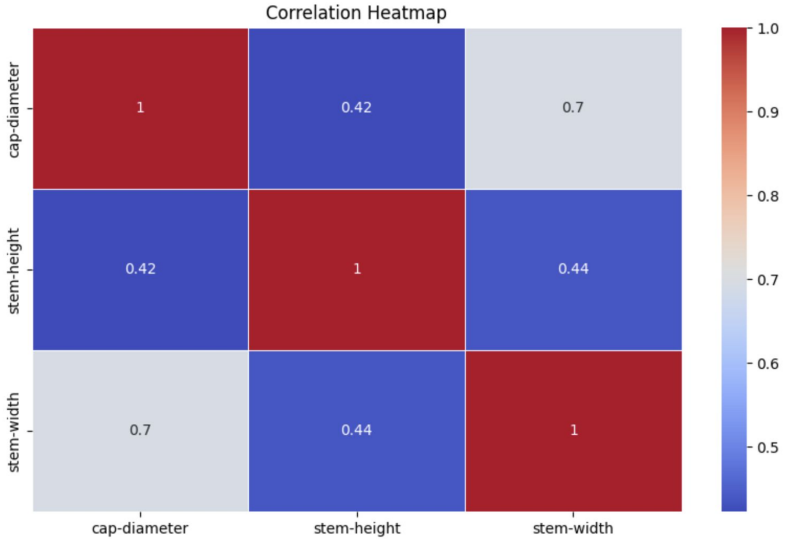
Approximate counts:

- **Poisonous (p):** 33,888
- **Edible (e):** 27,181

The dataset is **not perfectly balanced**, but the difference is not extreme.

Exploratory Data Analysis

Correlation Analysis

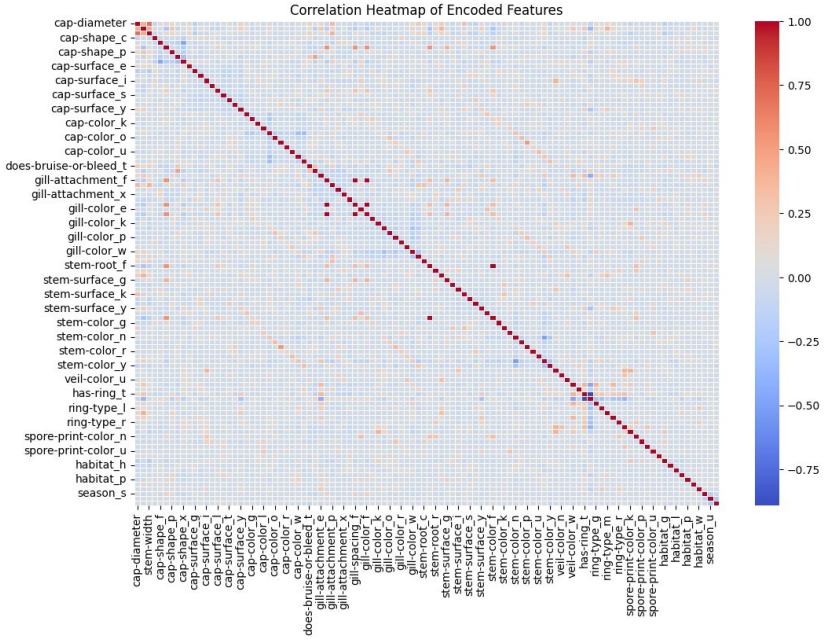


Correlation Heatmap of Numerical Features

- Correlations between **cap-diameter, stem-height, and stem-width**.
- **Cap-diameter and stem-width** have a **strong positive correlation (0.70)** – larger mushrooms tend to have wider stems.
- **Stem-height has moderate correlations** with both cap-diameter (0.42) and stem-width (0.44).
- These relationships suggest that **mushroom size-related features** could be useful in classification models.

Correlation Heatmap of Encoded Features

- Displays **correlation among one-hot encoded categorical variables**.
- Most features **do not show strong correlations**, indicating relatively independent categorical attributes.
- Some weak correlations suggest certain features may co-occur more frequently (e.g., **cap-surface and cap-color variations**).



Clustering–Factor Analysis

Choice of Dimension Reduction Method

Why PCA May Not Be a Suitable Dimensionality Reduction Method in This Case:

- **Low Explained Variance with First Few Components**

  Even after **10 components**, the cumulative explained variance is only around **60%**

- **Loss of Interpretability**

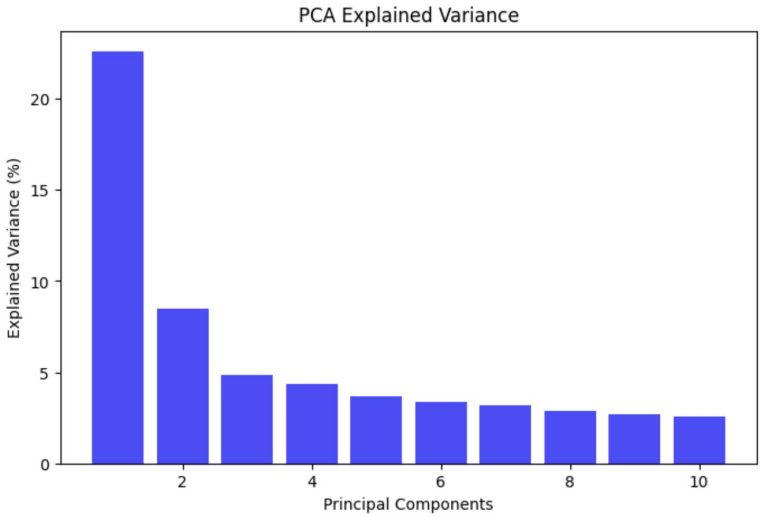
  Many categorical features (e.g., **cap shape, gill color, habitat**) are converted into numerical components, losing their biological meaning.

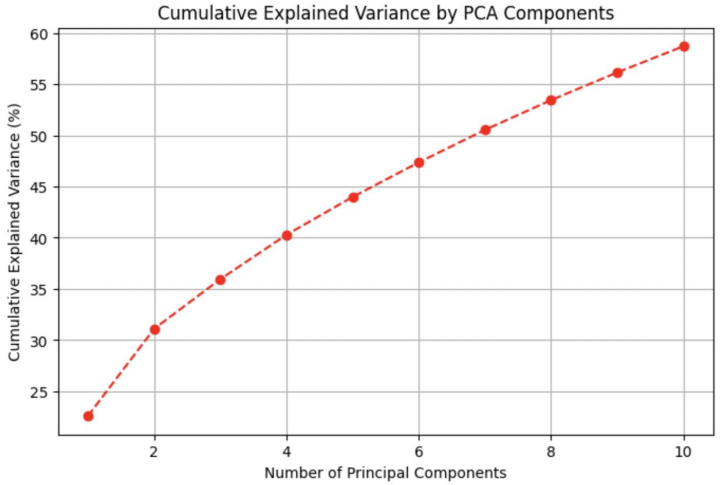
- **Dataset Contains Many Nominal Features**

  The mushroom dataset contains **mostly categorical variables**, making PCA less effective in reducing meaningful dimensions.

- **Non-Linear Relationships Exist in the Data**

  Methods like **t-SNE or UMAP** may better preserve non-linear relationships.





Clustering–Results & Interpretations

Identify natural groupings of mushrooms

t-SNE + DBSCAN Clustering

❌ **Over-fragmentation**

❌ The **-1 label (outliers)** is visible, but most points assigned to very tiny clusters

❌ **DBSCAN struggles**

❌**Extremely high Silhouette Score(0.826)**

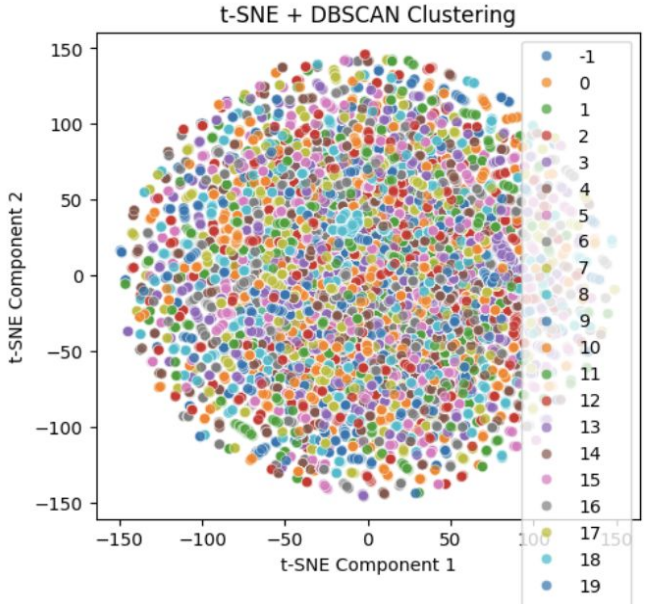
UMAP + DBSCAN Clustering

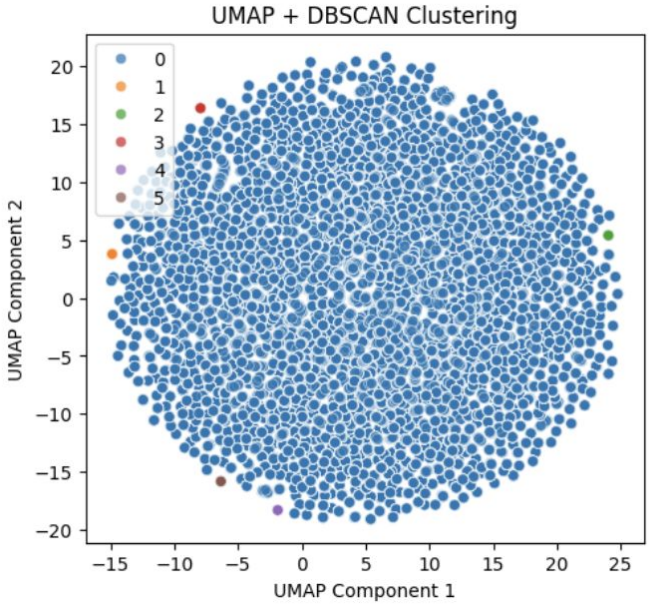
✅ **DBSCAN successfully detects outliers**

❌ **Almost all points** assigned to a **single cluster**

❌ Extremely low **Silhouette Score(-0.208)**

❌ **Outliers mistakenly identify as new cluster**





Clustering–Results & Interpretations

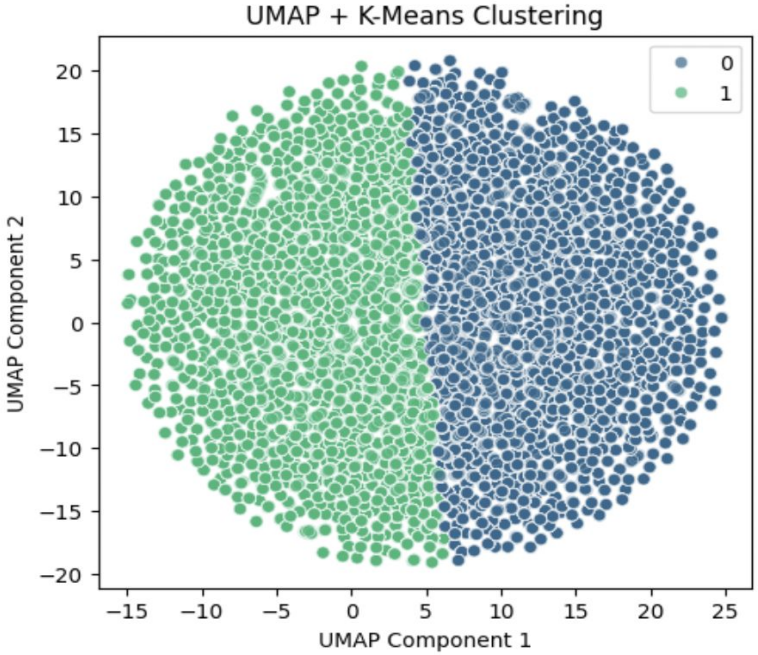
Identify natural groupings of mushrooms

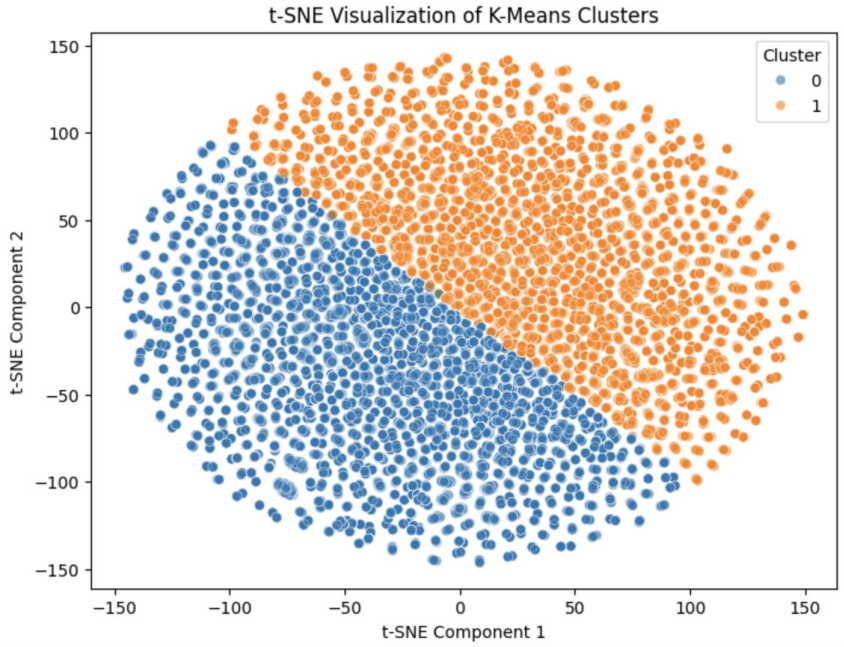
UMAP + K Means Clustering

- Silhouette Score: 0.327
- Davies-Bouldin Index: 1.241
- KNN Preservation Score: 0.100

t-SNE + K Means Clustering

- Silhouette Score: 0.330
- Davies-Bouldin Index: 1.236
- KNN Preservation Score: 0.101





Overall Results: Moderate clustering quality. Good separation of two clusters

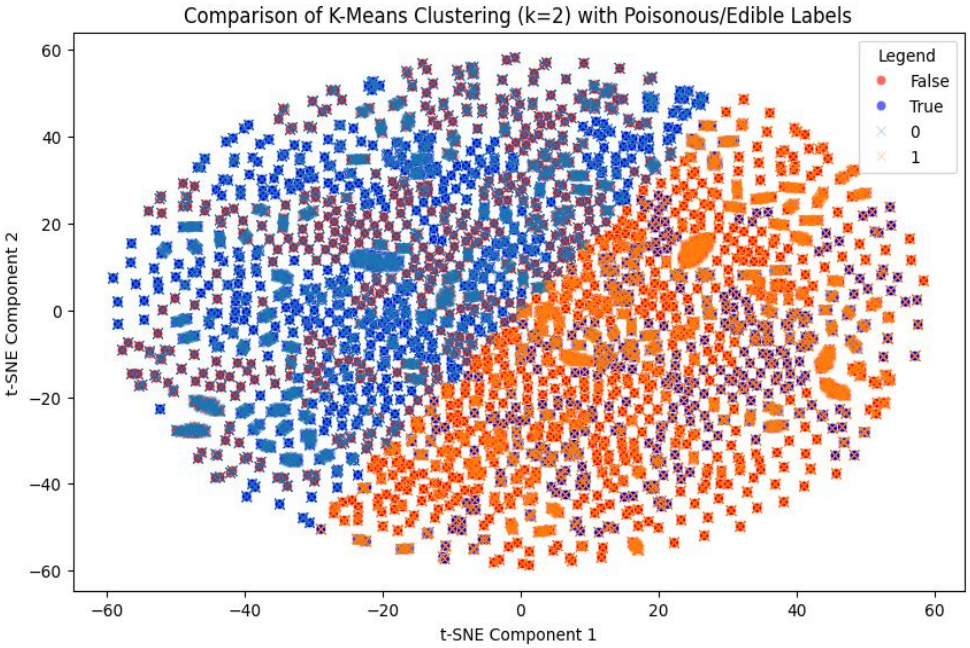
- tSNE somewhat better than UMAP in terms of silhouette score and Davies-Bouldin Index.

Clustering–Results & Interpretations

Identify natural groupings of mushrooms

K-Means with Poisonous/Edible Labels

- Suboptimal performance
- Overlappings in edible and poisonous mushrooms clustering



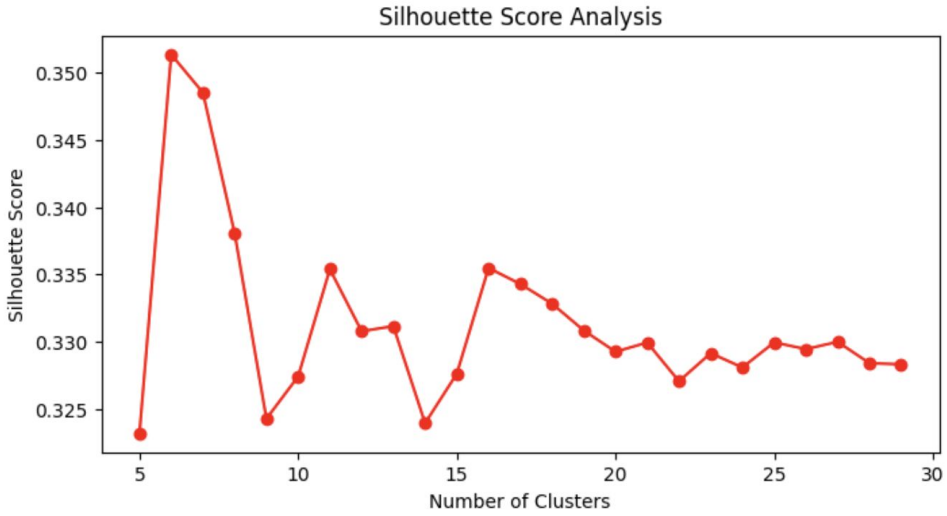
Possible Reasons for Overlap:

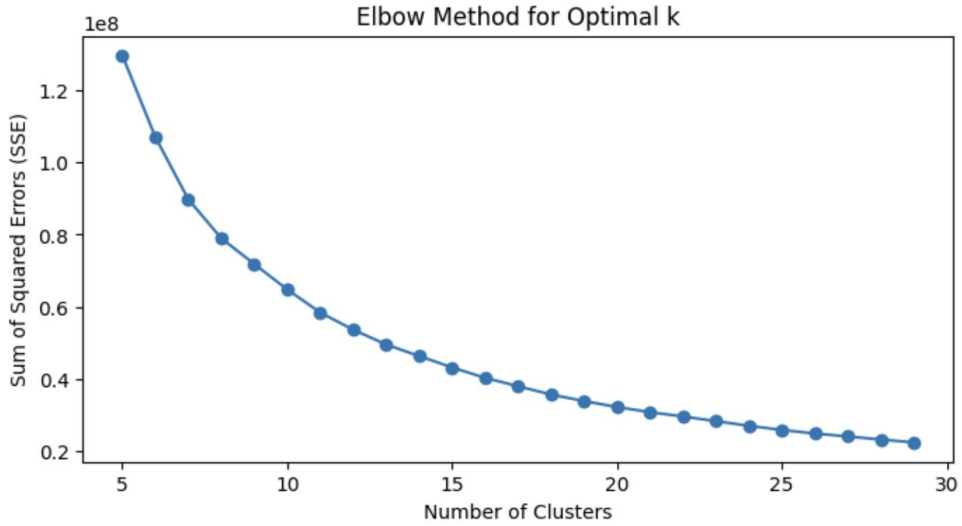
- **Feature Similarity** → Some edible and poisonous mushrooms share very similar **physical or chemical characteristics**.
- **t-SNE Projection Loss** → The visualization uses **t-SNE**, which **reduces dimensions** and might lose information.
- **Limitations of K-Means** → K-Means assumes **clusters are spherical**, but the actual distribution of mushrooms might not fit this assumption.

K-Means identifies some structure but is not a perfect classifier for identifying edible vs. poisonous mushrooms.

Clustering–Factor Analysis

Determine number of clustering





Elbow Method

- **Elbow Method:** SSE curve with **gradual decline** and no sharp elbow

Silhouette Score Analysis

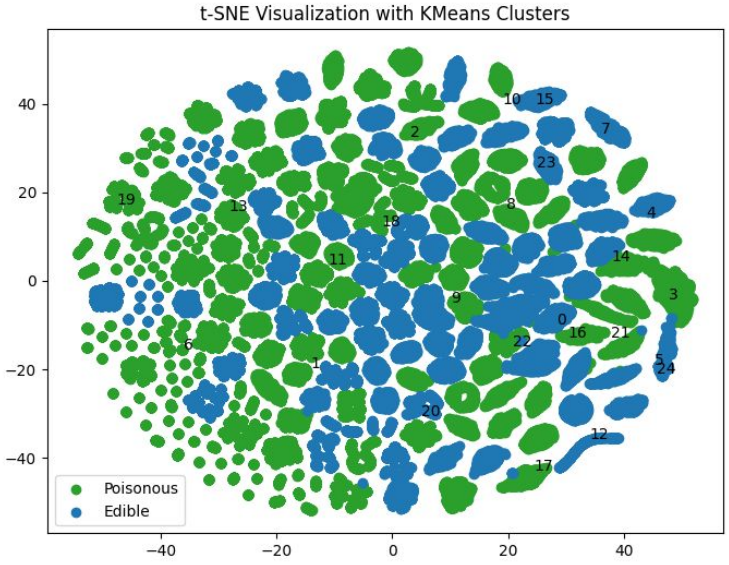
- **Silhouette Score:** fluctuates across different k values, with **peaks at lower k values**

Optimal K

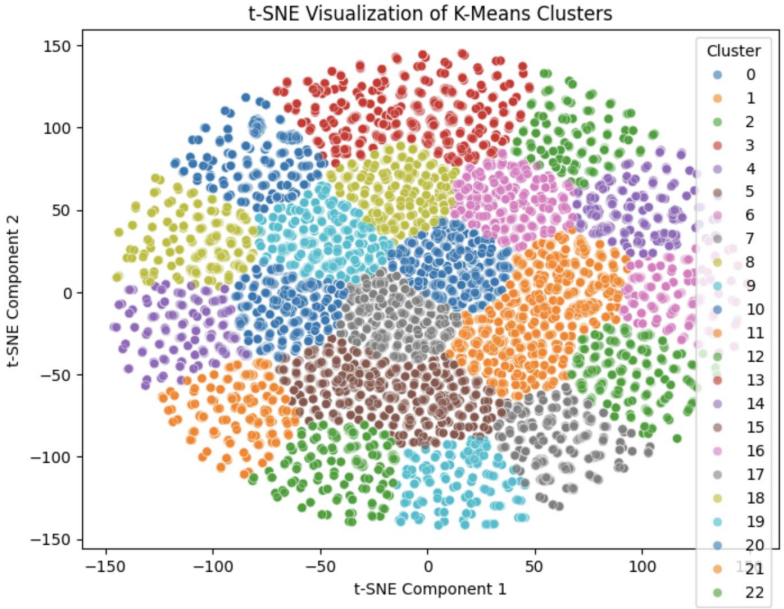
- Both stabilize **around k = 23**: coincides with number of mushroom species

Clustering – Conclusions and Recommendations

Mushrooms by species? Mushrooms by edibility?



- Natural Groupings of Mushroom: by **species** rather than edibility
  - improved Davies-Bouldin Index (0.804)
  - Mushroom edibility **not naturally separated** by feature similarity
  - Edibility **not the primary differentiator** or **too well-defined**
    - multi-layered edibility assessment?

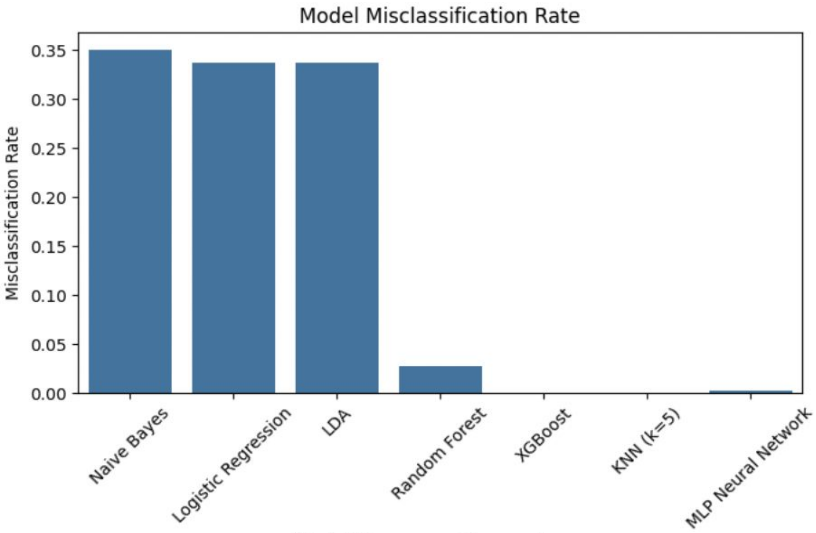


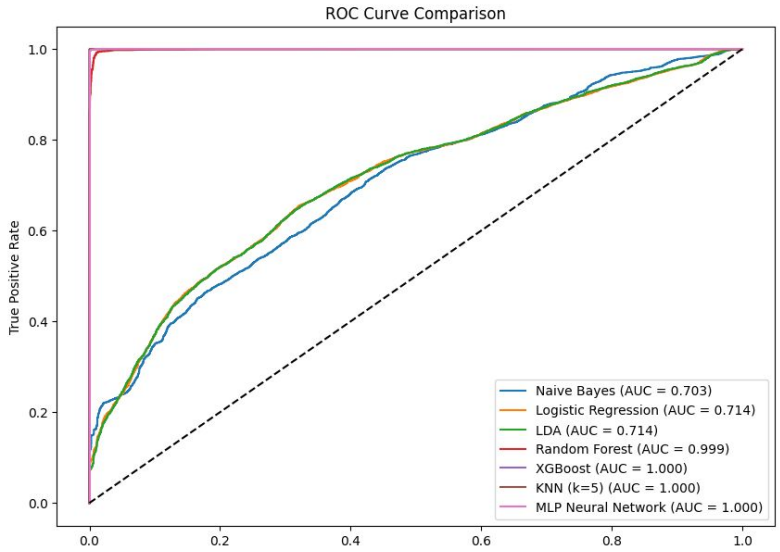
Clustering Improvement

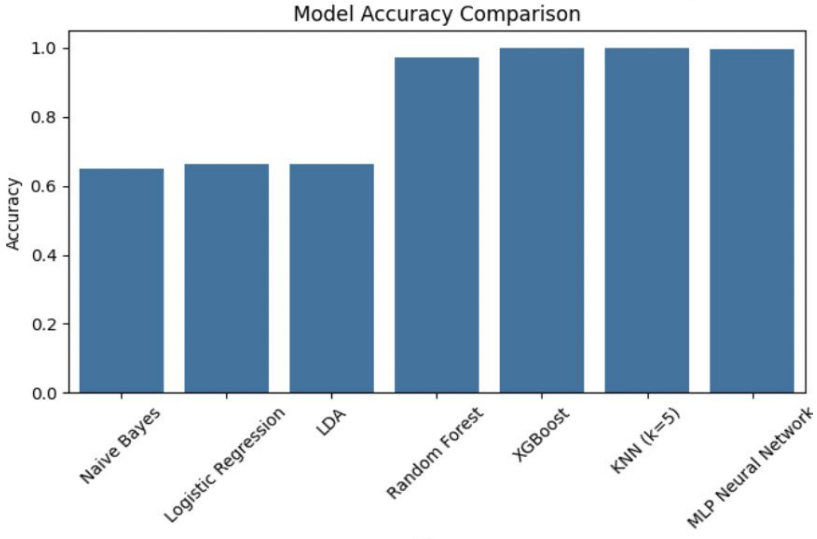
- different distance metrics
- models with different data shape assumptions

Classification–Results

Identify whether mushrooms are edible







Model	Accuracy	ROC-AUC	Misclassification Rate
Naïve Bayes	75%	0.838	0.253
Logistic Regression	66%	0.714	0.337
LDA	65%	0.715	0.349
Random Forest	98%	1.000	0.024
XGBoost	100%	1.000	0.000
KNN (k=5)	100%	1.000	0.001
MLP (Neural Network)	99.9%	1.000	0.003

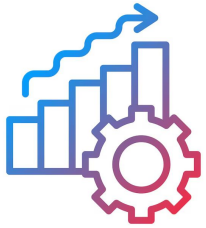
Classification–Interpretation & Insights

Identify whether mushrooms are edible

Best Models–Performance Metrics

Overall Performance Trends

- Naïve Bayes, Logistic Regression, and LDA had moderate performance, with accuracy between 65%-75%.
- Random Forest significantly outperformed basic models, achieving 98% accuracy.
- XGBoost, KNN (k=5), and MLP (Neural Network) achieved perfect or near-perfect classification with 100% accuracy.



Insight 1: Tree-Based Models Outperform Linear Models:

- **Random Forest and XGBoost** significantly outperform Naïve Bayes, LDA, and Logistic Regression.
- Linear models (Logistic Regression, LDA) struggle with feature interactions and class separation.

Insight 2: XGBoost and KNN (k=5) Achieve Perfect Classification:

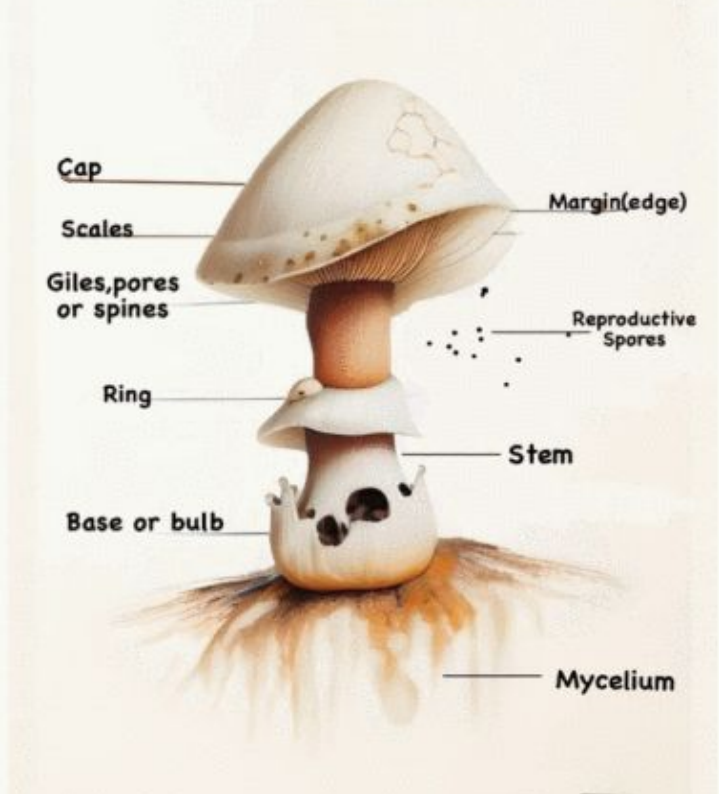
- XGBoost handles feature importance & interactions effectively.
- KNN works well in this dataset but may not scale efficiently.

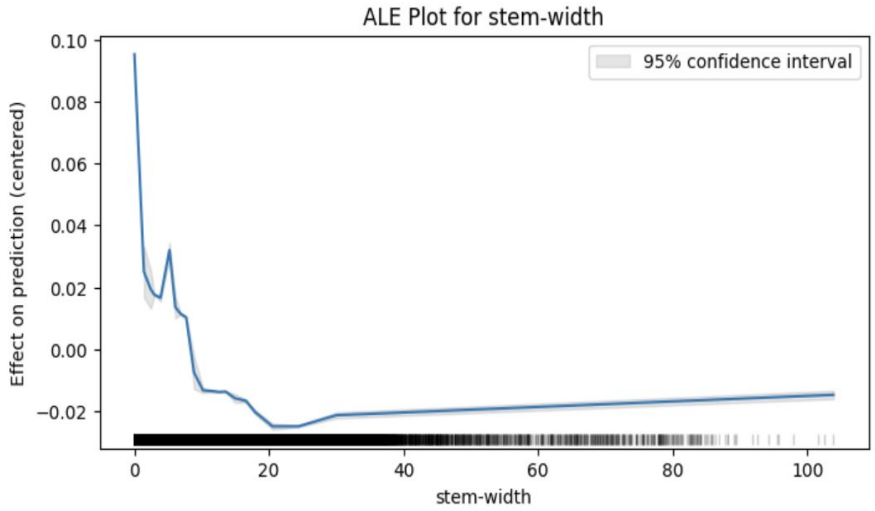
Classification–Models Explained

Identify important features for explaining mushroom edibility

Models Explained:

- Decision Tree + Logistic Regression
- Random Forest + ALE Plot





Key features affecting mushroom edibility:

- **Stem width (thin)**
- **Gill spacing (dense)**
- **Has ring**
- **Has bruise (red/orange)**

Future Steps and Business Impact

Highlight key strategies based on analysis and modelling result

Enhanced Food Safety & Compliance – Establishes metrics for classifying mushrooms, reducing misclassification risks and improving regulatory standards.
Research & Innovation – Provides a scalable, data-driven classification approach, reducing reliance on costly fieldwork and lab testing.

01

Deploy the Best Prediction Model in a Production Setting

- Convert best-performing models into an API using Flask/FastAPI.
- Deploy the model as a mushroom edibility identifier mobile app or website.

02

Consider Expanding to Multiclass Classification

- Current binary classification (edible vs. poisonous) could be expanded.
- Add multi-layered edibility assessment levels for real-world purposes (such as for medicine, for food).

03

Integrate External Data for Better Predictions

- Enhance model performance by adding environmental and contextual factors, such as: Geolocation (region-specific mushroom species), Weather conditions (humidity, temperature), Altitude and terrain (affects mushroom growth patterns)

04

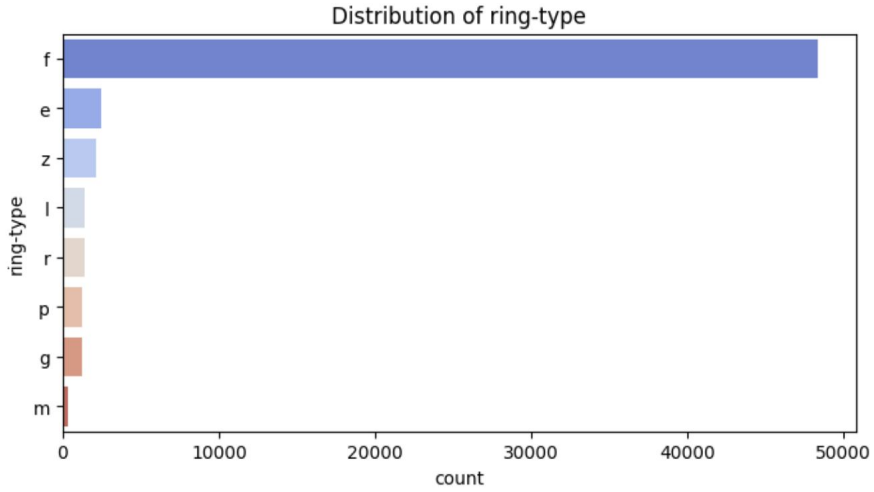
Employ mixture model for Better Feature Learning and Representation

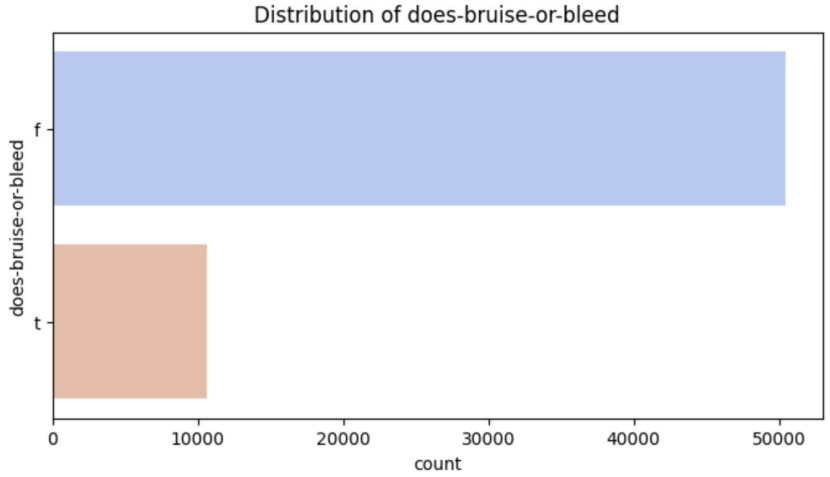
- Clusters mushrooms by species and then classify mushroom edibility to mine potential relationships between species and edibility (see appendix 5 for business application)

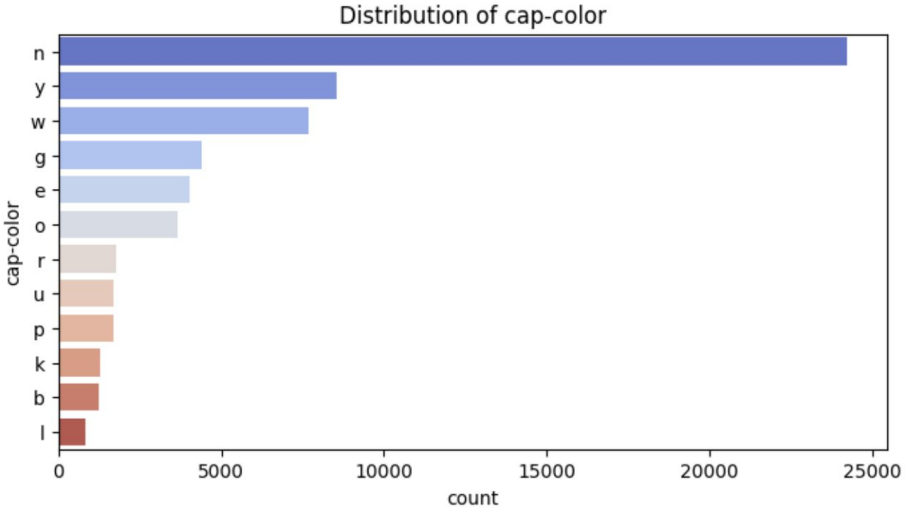
Thank you!

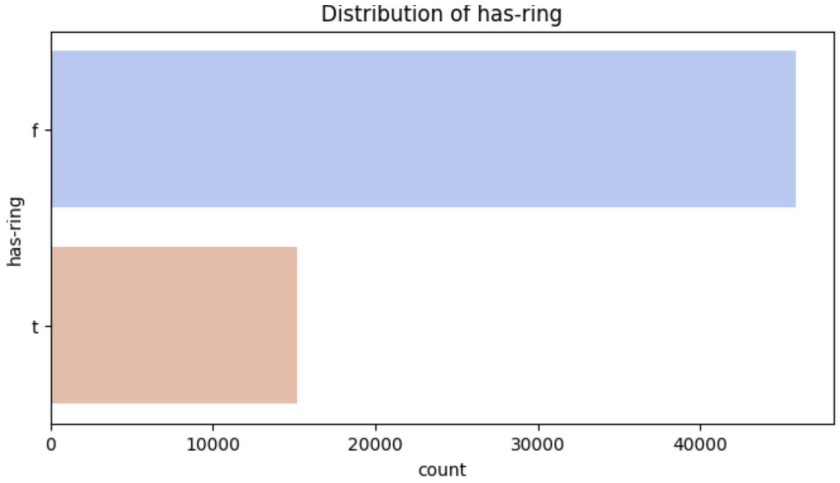
Appendix

Appendix 1: Most Important Categorical Variable Distribution

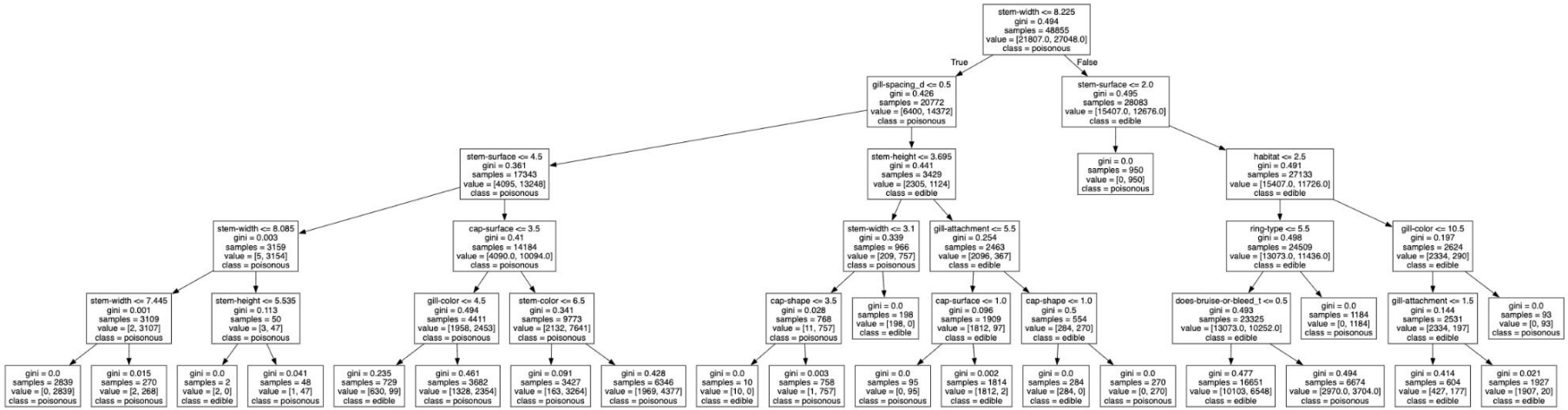


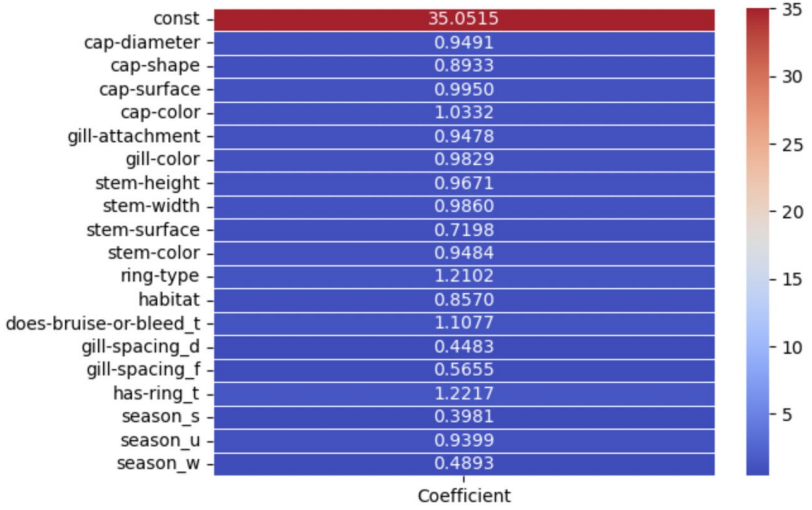


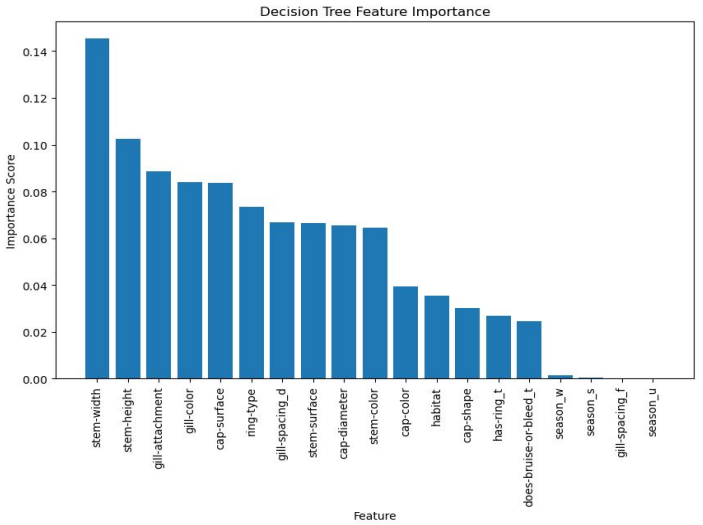




Appendix 2: Model Interpretation







Appendix 3: Future Study - Mixture Model

Technical Advantages

- Better feature learning and representation (**uncover hidden structure**)
- Reduce noise
- Better performance with **limited labeled data**
- **Higher flexibility** and **interpretability** for decision-making

Business Application Case

- **Poisonous** mushrooms with bioactive compounds of **high pharmaceutical value**
- Identify similar species with the same chemical structure
  - Lower research costs
  - Drug discovery opportunities