🍄 Machine Learning Project Report: Mushroom Edibility Classification

1. Project Goal and Overview

The main objective of this project was to determine whether a mushroom is **edible** or **poisonous** using data science and machine learning techniques. We started with a raw dataset of mushroom characteristics and followed a standard pipeline: **data cleaning**, **preprocessing**, **exploratory analysis**, **dimensionality reduction (PCA)**, **unsupervised learning (K-Means)**, and finally, **supervised classification (Random Forest)**.

2. Preparing the Data (The Cleanup Phase)

Before any machine learning can happen, the data must be clean and organized.

Competency	Action Taken	Why it Matters
Data Cleaning	**Missing values** (represented by '?') were filled in with the most common value (mode). A feature with only **one unique value** (the 'p.2' feature) was **removed** as it provided no information.	Ensures the data is complete and every feature contributes unique information.
Preprocessing	All categorical codes (e.g., 'k' for black, 'e' for edible) were converted into **numerical columns** using **One-Hot Encoding** and **Label Encoding**.	Machine learning models only understand numbers. This step expanded our initial 22 columns to **116 numerical features**.
Data Split	The data was split into a **Training Set (67%)** and a **Testing Set (33%)**.	The model trains on the Training Set and is tested on unseen data (the Testing Set) to ensure it performs well in the real world.

3. Exploring Data Structure and Complexity

With 116 features, the data is complex. We used **Principal Component Analysis (PCA)** to simplify and visualize it.

A. Dimensionality Reduction (PCA)

PCA helps us find the most important combinations of features that capture the most variation in the data.

- **Visualization:** We reduced the data to just **two components** to create a 2D scatter plot, colored by the target (edible/poisonous).

    - **Finding:** The plot showed large regions where the two mushroom types were perfectly separated, suggesting the classification task would be easy.
- **Information Retention:** We checked how many components are actually needed to keep
- 95%
- **95%** of the original information.
    - **Finding:** We still needed **109 out of the 116 components**. This means while the task is simple for a classifier, the features themselves are largely **non-redundant**. The complexity (116 features) is justified.

B. Unsupervised Clustering (K-Means)

We then used **K-Means clustering** (unsupervised learning) to see how the data naturally groups itself, without using the 'edible/poisonous' labels.

- **Optimal Clusters:** The **Elbow Method** confirmed that
- $K=2$
- $\mathbf{K=2}$ clusters were most appropriate (matching our two target classes).
- **Finding:** When plotting the clusters against the true labels, the K-Means groups were highly **mixed** (
- $\approx 50/50$
- $\approx 50/50$ split) .
- **Implication:** This shows that the factors that make a mushroom edible/poisonous **do not perfectly align** with the factors that make the mushroom features *look* similar. This is where a supervised model excels.

---

4. Supervised Classification and Final Results

We chose the **Random Forest** model—an Ensemble Method known for accuracy—to classify the mushrooms.

Competency	Model Used	Result
Classification Model	**Random Forest Classifier** (trained on standardized features)	**Test Accuracy:** 1.0000 **1.0000 (100%)**
Performance Check	Retrained Random Forest using the 109 **109** PCA components.	**Test Accuracy:** 1.0000 **1.0000 (100%)**

Conclusion

The **Random Forest** model achieved **perfect accuracy** (

100%

100%) in distinguishing between edible and poisonous mushrooms. This confirms that even though the dataset has many features, there is a **strong, clean signal** within the data that allows for flawless separation. The model is highly reliable for this classification task.

All figures and reports produced during this analysis are saved in the `results/figures/` and `results/reports/` directories for full reproducibility.