[Mushroom]

[Tobias Manthey, Sander Birkhol, Andrija Bojic], [01.11.205]

1: DESCRIBE THE PROBLEM

Scope

The goal of this project is to develop a machine learning–based web application that predicts whether a mushroom is edible or poisonous based on its physical characteristics. Users select attributes such as cap shape, color, odor, gill size, stalk features, and habitat. The system outputs a prediction of either *edible* or *poisonous*.

Machine learning is a promising solution because mushroom classification involves subtle patterns across many categorical features. These patterns can be difficult for people without botanical expertise to evaluate manually. Machine learning models, however, can learn these relationships from data and provide instant predictions.

Today, identifying mushroom safety generally requires expert knowledge, field guides, or chemical testing. Manual identification is slow, requires domain knowledge, and is prone to dangerous human error. A machine learning solution offers speed, accessibility, and consistency.

Business objective:
The product aims to increase accessibility to mushroom classification tools for educational and research use. It provides value by simplifying mushroom identification and enabling quick assessments. While not intended as a real-life safety tool, it demonstrates how ML can support categorical decision-making in biological datasets.

User group:
Students, hobbyists, and individuals learning about ML classification and mushroom attributes.

Success criteria (business metric):

- High prediction accuracy (≥ 95%)
- Good user experience and ease of use (web-based interface)

Metrics

The main ML metric is **accuracy**, measuring the share of correct predictions on unseen data. The RandomForest model achieves approximately **99% accuracy**, which exceeds the minimum acceptable performance target of 95%.

Latency and throughput are also relevant, as the prediction occurs in real time in a web application. Prediction time is below 100ms, ensuring a responsive UI.

From a business perspective, success is defined by:

- Model accuracy ≥ 95%
- Users being able to complete predictions quickly and without technical knowledge

---

2: DATA

The model uses the **Agaricus–Lepiota Mushroom Dataset**, originally from the UCI Machine Learning Repository (also available via Kaggle). The dataset contains **8124 rows and 22 categorical features**, each representing characteristics of mushrooms.

Each record includes properties such as:

- Cap shape, color, surface, gill attachment

The label represents mushroom class:

- **e = edible**
- **p = poisonous**

All features are categorical. No preprocessing beyond encoding was needed. The data was cleaned to ensure that missing values (represented as ? in the stalk-root column) were preserved as a valid category.

Labels:
Ground truth labels were provided in the dataset, eliminating manual labeling work and ensuring consistent annotation.

Ethical considerations:
Because the model predicts mushroom edibility, there is a safety risk if someone misuses the

application for real-world mushroom foraging. Therefore, a explicit disclaimer is included to prevent misuse. No personal or sensitive user data is collected.

Feature representation and preprocessing:

- One-hot encoding / label encoding was used to convert categorical features into numeric values
- Data was loaded into a Pandas DataFrame for training
- No scaling was required due to the categorical nature of the dataset

---

3: MODELING

The selected model is a **RandomForestClassifier**, chosen because:

- It performs well with categorical data (after encoding)
- It handles nonlinear decision boundaries
- It provides feature importance insights

A simple baseline classifier (predicting the majority class every time) was used to assess baseline performance (~50%). The RandomForest model significantly outperformed this, achieving **~99% test accuracy**, demonstrating that the dataset has strong patterns separating edible and poisonous mushrooms.

Model evaluation included:

- Train/test split
- Accuracy score calculation
- Manual inspection of prediction errors (very few misclassifications observed)

Feature importance inspection showed that features such as *odor*, *spore print color*, and *gill size* played the largest role in predictions. Exploring these insights guided minor tuning of the model to ensure generalization.

4: DEPLOYMENT

The final system is deployed as a **Flask web application**.
https://mushroom-detection-ml.onrender.com/

Users interact with a graphical interface (`index.html`) that presents dropdown menus for each mushroom characteristic. When the user submits the form, the values are sent to the Flask backend (`app.py`). The machine learning model and label encoder (stored as `.pkl` files) generate a prediction and return one of two outputs:

- "This mushroom is edible!"
- "This mushroom is poisonous!"

The web interface is styled using a custom CSS file (`style.css`).

5: REFERENCES

- Dua, D. & Graff, C. (2017). UCI Machine Learning Repository. Mushroom Dataset. https://archive.ics.uci.edu/ml/datasets/Mushroom
- Kaggle. Mushroom Classification Dataset (Agaricus–Lepiota). https://www.kaggle.com/uciml/mushroom-classification
- Scikit-learn documentation (RandomForestClassifier). https://scikit-learn.org/