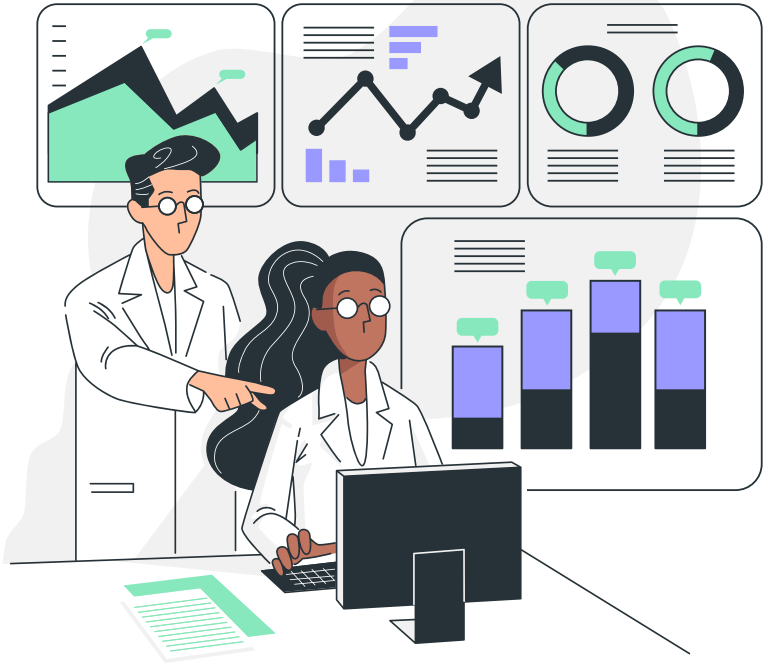

Binary Prediction of Poisonous Mushrooms

ARTIFICIAL INTELLIGENCE

Bruno Drumond up201202666
Tomás Sucena Lopes up202108701

Problem Specification

GOAL

Develop a machine learning model to predict whether a mushroom is **poisonous** or **edible** based on its physical characteristics (e.g., cap shape, veil type, gill color).

WHY THIS MATTERS

- Mushroom foraging is **risky** due to toxic species.
- A reliable model can **support education, safety applications, and preliminary field classification tools**.

DATASET DESCRIPTION

- Based on the **UCI Mushroom Dataset**
- Contains over **3 million samples** and **21 features**
- **Categorical and quantitative data**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3116945 entries, 0 to 3116944
Data columns (total 22 columns):
 #   Column                Dtype
---  ------                -----
 0   id                    int64
 1   class                 object
 2   cap-diameter          float64
 3   cap-shape             object
 4   cap-surface           object
 5   cap-color             object
 6   does-bruise-or-bleed  object
 7   gill-attachment       object
 8   gill-spacing          object
 9   gill-color            object
 10  stem-height           float64
 11  stem-width            float64
 12  stem-root             object
 13  stem-surface          object
 14  stem-color            object
 15  veil-type             object
 16  veil-color            object
 17  has-ring              object
 18  ring-type             object
 19  spore-print-color     object
 20  habitat               object
 21  season                object
dtypes: float64(3), int64(1), object(18)
memory usage: 523.2+ MB
```

Related Work and References

EXISTING RESEARCH

- Many studies have used the **UCI Mushroom Dataset** for binary classification challenges.
- Most approaches explore **Decision Trees**, **Random Forests**, or **Naive Bayes** due to categorical features.
- The existing work confirms that a simple model with categorical preprocessing can yield high accuracy.

KEY REFERENCES

- https://www.kaggle.com/code/annastasy/ps4e8-data-cleaning-and-eda-of-mushrooms
- https://github.com/Kolwankar-Siddhiraj/MushroomClassificationProjectML
- https://ai.plainenglish.io/mushroom-classification-using-machine-learning-with-deployment-using-fastapi-16ff80bc4cef
- https://medium.com/analytics-vidhya/mushroom-classification-using-different-classifiers-aa338c1cd0ff

Tools and Algorithms

Language

Python (using the Jupyter Notebook) since it offers robust tools and community support

Libraries

Pandas/Numpy: Data manipulation and preprocessing

Matplotlib/Seaborn: Visualization and data inspection

Scikit-learn: Machine learning models and performance metrics

Algorithms Used

Decision Tree: Easy to interpret and handles categorical inputs well

Random Forest: Reduces overfitting and improves accuracy and robustness

Logistic Regression: Interpretable model and establishes a reliable performance baseline

Checkpoint Progress

Data Loading
Read from compressed .zip file using pandas

Data Cleaning
Removed ***id*** column and dropped duplicate rows to reduce noise

Data Preprocessing
(view next slide)

Model Training
Used ***DecisionTreeClassifier*** from ***sklearn***

Initial Evaluation
Calculated accuracy score on test set

To be continued...

The decision tree model was successfully trained with initial promising accuracy!

Data Preprocessing

IMPUTING MISSING VALUES

- **Quantitative Data**
  - Assessed the skewness of each column to determine if it was more appropriate to impute with the **average** or the **median**.
  - As all columns were right-skewed, the **median** was chosen.
- **Qualitative Data**
  - Computed the percentage of missing values from each column.
  - Replaced missing values with the **mode** if the percentage was low, otherwise with a new value - *Unspecified*.

HANDLING OUTLIERS

- **Quantitative Data**
  - Removed all rows with values outside the range $[Q_{0.10}, Q_{0.90}]$.