Mushroom Mystery: Predicting Edible vs. Poisonous Varieties Using Machine Learning Models

Goal: To predict whether a mushroom is edible or poisonous based on its physical characteristics.

Evaluation: Submissions are evaluated using the Matthews correlation coefficient (MCC).

Dataset Description : The dataset for this Kaggle competition (both train and test) was generated from a deep learning model trained on the UCI Mushroom dataset. Feature distributions are close to, but not exactly the same, as the original. Feel free to use the original dataset as part of this competition, both to explore differences as well as to see whether incorporating the original in training improves model performance.

Meta Data

Feature Name	Description
id	Unique identifier for each mushroom instance
class	Binary target variable (`e` for edible, `p` for poisonous)
cap-diameter	Diameter of the mushroom cap (numeric)
cap-shape	Shape of the mushroom cap (categorical)
cap-surface	Surface texture of the mushroom cap (categorical)
cap-color	Color of the mushroom cap (categorical)
does-bruise-or-bleed	Indicates whether the mushroom bruises or bleeds when handled (categorical)
gill-attachment	Attachment type of the gills to the stem (categorical)
gill-spacing	Spacing between the gills (categorical)
gill-color	Color of the gills (categorical)
stem-height	Height of the mushroom stem (numeric)
stem-width	Width of the mushroom stem (numeric)
stem-root	Root structure of the mushroom stem (categorical)
stem-surface	Surface texture of the mushroom stem (categorical)
stem-color	Color of the mushroom stem (categorical)
veil-type	Type of the veil covering the mushroom (categorical)
veil-color	Color of the veil (categorical)
has-ring	Indicates if the mushroom has a ring around the stem (categorical)
ring-type	Type of ring present around the mushroom stem (categorical)
spore-print-color	Color of the spore print (categorical)
habitat	Natural habitat of the mushroom (categorical)
season	Season in which the mushroom grows (categorical)

Note: Unlike many previous Tabular Playground datasets, data artifacts *have not* been cleaned up. There are categorical values in the dataset that are not found in the original. It is up to the competitors how to handle this.

Input Files

- **train.csv** - the training dataset; class is the binary target (either e or p)
- test.csv - the test dataset; your objective is to predict target class for each row