CSM148 Project 3 Specification

Due date: Friday, March 3 at 11:59PM PST
Instructions: All work must be completed individually. Submit a PDF of your Jupyter notebook with all code blocks' output visible and a separate PDF for your report to GradeScope.

1 Overview

You are exploring the wilderness of *Mushroomia*, a land populated by a plethora of diverse fauna and flora. In particular, *Mushroomia* is known for its unparalleled variety in mushrooms. However, not all the mushrooms in *Mushroomia* are edible. As you make your way through *Mushroomia*, you would like to know which mushrooms are edible, in order to forage for supplies for your daily mushroom soup.

You have access to:

- *Shroomster Pro Max* ™ - a state of the art data collection device, developed by *Mushroomia*, that allows you to collect various data points about any mushroom you encounter in the wild
- *The National Archives on Mushrooms* - a dataset collected over the years by the government of *Mushroomia*

To address this problem, you decide to use the skills you learnt in CSM148 and train machine learning models on the *The National Archives on Mushrooms* in order to use your *Shroomster Pro Max* ™ to determine whether the mushrooms you encounter on your adventure can be added to your daily mushroom soup.

This project will be more unstructured than the previous two projects in order to allow you to experience how data science problems are solved in practice. There are two parts to this project: a Jupyter Notebook with your code (where you explore, visualize, process your data and train machine learning models) and a report (where you explain the various choices you make in your implementation and analyze the final performance of your models).

2 Dataset

Labels: One binary class divided in poisonous=p (to be encoded as class 0) and edible=e (to be encoded as class 1)

Features: These are the features that the *Shroomster Pro Max ™* can determine for mushrooms in the wild

1. *cap-diameter:* float number in cm
2. *cap-shape:* bell=b, conical=c, convex=x, flat=f, sunken=s, spherical=p, others=o
3. *cap-surface:* fibrous=i, grooves=g, scaly=y, smooth=s, shiny=h, leathery=l, silky=k, sticky=t, wrinkled=w, fleshy=e
4. *cap-color:* brown=n, buff=b, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y, blue=l, orange=o, black=k
5. *does-bruise-bleed:* bruises-or-bleeding=t,no=f
6. *gill-attachment:* adnate=a, adnexed=x, decurrent=d, free=e, sinuate=s, pores=p, none=f, unknown=?
7. *gill-spacing:* close=c, distant=d, none=f
8. *gill-color:* see cap-color + none=f
9. *stem-height:* float number in cm
10. *stem-width:* float number in mm
11. *stem-root:* bulbous=b, swollen=s, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r
12. *stem-surface:* see cap-surface + none=f
13. *stem-color:* see cap-color + none=f
14. *veil-type:* partial=p, universal=u
15. *veil-color:* see cap-color + none=f
16. *has-ring:* ring=t, none=f
17. *ring-type:* cobwebby=c, evanescent=e, flaring=r, grooved=g, large=l, pendant=p, sheathing=s, zone=z, scaly=y, movable=m, none=f, unknown=?
18. *spore-print-color:* see cap color
19. *habitat:* grasses=g, leaves=l, meadows=m, paths=p, heaths=h, urban=u, waste=w, woods=d
20. *season:* spring=s, summer=u, autumn=a, winter=w

Data Files: Train Set has 50213 rows, Test Set has 10856 rows.

`mushroom_train.csv, mushroom_test.csv`

3 Jupyter Notebook: Coding Requirements (50 pts)

Please organize your code using the outline (i.e. section titles) provided in the starter notebook for this project and upload your notebook with all code blocks' output visible as a PDF.

1. **Loading and Viewing Data (2 pts)**
2. **Splitting Data into Features and Labels (2 pts)**
3. **Data Exploration and Visualization (8 pts)**
4. **Data Processing (10 pts)**
5. **Data Augmentation (Creating New Features) (4 pts)**
6. **Statistical Hypothesis Testing (6 pts)**
7. **Dimensionality Reduction using PCA (2 pts)**
8. **Experiment with any 2 other models (Non-Ensemble) (4 pts)**
9. **Experiment with 1 Ensemble Method (2 pts)**
10. **Cross-Validation & Hyperparameter Tuning for All 3 Models (6 pts)**
11. **Report Final Results (4 pts)**

4 Report Requirements (50 pts)

Please submit the report as a separate PDF file.

1. **Introduction: (3 pts)** a brief overview of the problem, your methodology and your final results
2. **Methodology:**
   (a) **Data Loading, Splitting, Exploration and Visualization: (6 pts)** what did you explore / visualize? why did you explore / visualize it in this way? what did you learn from this about the data?
   (b) **Data Pre-Processing: (4 pts)** how did you pre-process your data? why did you pre-process your data in this way?
   (c) **Data Augmentation: (4 pts)** what additional features did you create? why do you think they are useful?
   (d) **Statistical Hypothesis Testing: (10 pts)** what relationships between variables did you investigate? were the results statistically significant? what did you learn from this about the data?
   (e) **Models of your choice (2 distinct models): (6 pts)** what models did you choose to implement? why are they appropriate for the problem?
   (f) **Ensemble Method (1 ensemble method): (3 pts)** what ensemble method did you use? why is it a good fit for this problem?
   (g) **Hyper-parameter Tuning: (3 pts)** how did you tune the hyperparameter for each model? what were the best parameters you found?
3. **Results: (4 pts)** Compare the performance of each of the models you implemented using the standard evaluation metrics. Discuss the cross-validation strategy you used and why you chose this strategy.
4. **Conclusion: (7 pts)** Decide which model you would use on your adventure through *Mushroomia*. Explain why you chose this model based on the results. Discuss any limitations of your project (e.g. limitations of model, data analysis and visualization, concerns with the raw data in *National Archives on Mushrooms*).