Assignment – Preprocessing Data for scikit-learn

Very often, we're tasked with taking data in one form and transforming it for easier downstream analysis. In this assignment, you'll use what you've learned in the course to prepare data for predictive analysis in Project 4.

Mushrooms Dataset. A famous—if slightly moldy—dataset about mushrooms can be found in the UCI repository here: https://archive.ics.uci.edu/ml/datasets/Mushroom. The fact that this is such a well-known dataset in the data science community has made it a good dataset to use for comparative benchmarking. For example, if someone was working to build a better decision tree algorithm (or other predictive classifier) to analyze categorical data, this dataset could be useful. In Project 4, we'll use `scikit-learn` to answer the question, "Which other attribute or attributes are the best predictors of whether a particular mushroom is poisonous or edible?"

Your assignment is to

- First study the dataset and the associated description of the data (i.e. "data dictionary"). You may need to look around a bit, but it's there!
- Create a `pandas DataFrame` with a subset of the columns in the dataset. You should include the column that indicates edible or poisonous, the column that includes odor, and at least one other column of your choosing.
- Add meaningful names for each column.
- Replace the codes used in the data with numeric values—for example, in the first "target" column, "e" might become 0 and "p" might become 1. This is because your downstream processing in Project 4 using `scikit-learn` requires that values be stored as numerics.
- Perform exploratory data analysis: show the distribution of data for each of the columns you selected, and show scatterplots for edible/poisonous vs. odor as well as the other column that you selected.
- Include some text describing your preliminary conclusions about whether either of the other columns could be helpful in predicting if a specific mushroom is edible or poisonous.

Your deliverable is a Jupyter Notebook that performs these transformation and exploratory data analysis tasks.

If you are working in a group, you also have the option of replacing the mushroom dataset in the assignment with a different data set that your group members might find more interesting.

You should post the Jupyter Notebook (.ipynb) file in your GitHub repository, and provide the appropriate URL to your GitHub repository in your assignment link. You should also have the original data file accessible through your code—for example, read directly from the UCI repository or stored in a GitHub repository.