Die or not?

This dataset includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family Mushroom drawn from The Audubon Society Field Guide to North American Mushrooms (1981). Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one. The Guide clearly states that there is no simple rule for determining the edibility of a mushroom; no rule like "leaflets three, let it be'' for Poisonous Oak and Ivy.

Mushroom hunting is enjoying new peaks in popularity. At the end of this project I will figure it out which features spell certain death and which are most palatable in this dataset of mushroom characteristics.

For completing this project, first I will study the dataset using various tools that scikit-learn have (heatmap, feature selection, learning curve). Base of the evaluation I will decide which features have the most impact of the final result. Next, I will try to minimize the noise from the dataset by removing the features with least impact of the result. Next, I will decode the dataset for a better accuracy result, and then I will apply multiple models to the same dataset for large evaluation.

At the end of this project I will be able to tell what type of machine learning models perform best on this dataset. And which features are most indicative of a poisonous mushroom.

For this project I will use 5 of the most well know machine learning models.

Logistic Regression

Gaussian Naïve Bayes

Support Vector Machine

Random Forest Classifier

Decision Tree