Deliverable 1

1. Choice of Dataset:

I chose the mushroom classification dataset because I love mushrooms. In fact, I wanted to become a mycologist at some point when I was younger.

2.Methodology:

a. The dataset I chose seems very easy to work with. The mushrooms have 23 different features, one of them being “class”. Class refers to whether or not the mushroom is edible or poisonous, and will be my label. The other 22 features all have non numerical values, so I will have to convert those values into numbers. I Should also check the correlation between those other features and the “class” feature, that way I could eliminate features that poorly correlate with my label. I’ll also have to split the data into training and testing sets.

b. This is a classification problem; the mushrooms are either edible or poisonous. I am probably either going to use an SVM classifier, or a random forest classifier. They are beginner friendly and compatible with my dataset without too much data pre-processing.

c. Because this is a classification problem, a confusion matrix and a simple accuracy test should suffice. I hope to reach an accuracy that is close to 100%, it seems very achievable by looking at other code examples on Kaggle.

3. Application:

My web page will allow the user to input all of the mushrooms relevant features (text selection type input), and will then output whether or not it is poisonous or not.