Deliverable 2

Saiyara Islam

1. The objective was to classify mushrooms as edible or poisonous based on some visual features and spore print color (not habitat or population.) This involves text-based work only and uses random forest classifier (instead of a single decision tree.)
2. The dataset was retrieved from Kaggle. There are 20 features and 8124 samples. I turned the dataset from a text-based one to a numerical using LableEncoder. I didn't include habitat or population as features because I wanted to focus on features that are immediately visible to the average person.
3. Pandas, numpy, scikit-learn preprocessing and ensemble were used. The model uses random forest classifier (I am not sure about the details of the algorithm used by scikit-learn. The ratio of the number of samples in training, test and validation is 3:1:1. This was chosen based on a source online. https://www.malicksarr.com/split-train-test-validation-python/#:~:text=Split%20the%20dataset,use%20as%20the%20test%20set.
   Testing and validation incurred a problem unfortunately. When trying to find the score of the model I'm getting 1.0 which is very strange. I am not sure if I split the dataset properly.
4. The results weren't coming right. I'm getting a score of 1.0 for test and validation sets.
5. I need to fix this problem. After that I might adjust some hyperparameters (like the number of random forests used by the algorithm) to see if it yields a better prediction.