**Question 2.** *When evaluating supervised NB across the four different datasets, you will observe some variation in effectiveness (e.g. Accuracy). Explain what causes this variation. Describe and explain any particularly surprising results.*

The average accuracy of five runs was recorded (using a hold-out strategy with an 80-20 split) for each dataset:

Breast cancer: 71.93%

Car: 85.61%

Hypothyroid: 95.36%

Mushroom: 99.50%

The breast cancer performs by far the worst, whereas the mushroom has nearly 100% accuracy. The most obvious difference between these datasets is their size; the breast cancer data set has only 286 records, the car dataset has 1,728, the hypothyroid dataset has 3,163, and the mushroom dataset has 8,124 records. This correlates with the increase in accuracy; the more records a dataset has, the more accurate the probabilities will be, and therefore the more accurate the classifier will be. The accuracy increases are smaller on each step up in the number of instances, indicating diminishing returns in increasing the size of the dataset.

In a special case, the hypothyroid dataset was lopsided, especially when compared to the more accurate mushroom dataset. In the mushroom dataset, 51.8% of the dataset is of class ‘e’, and the rest ‘p’, whereas in the hypothyroid dataset, 95.22% of the dataset is of class ‘negative’, and the ‘hypothyroid’. Thus, the classifier would just predict ‘negative’ for every instance in the hypothyroid set, making it less accurate.

**Question 3.** *Evaluating the model on the same data that we use to train the model is considered to be a major mistake in Machine Learning. Implement a hold–out (hint: check out numpy.shuffle()) or cross–validation evaluation strategy. How does your estimate of Accuracy change, compared to testing on the training data? Explain why. (The result might surprise you!)*

This question was answered in the context of the breast cancer data set and using the supervised classifier. An 80-20 holdout strategy was implemented. Using the holdout strategy, the average accuracy over five runs was, to two decimal places, 71.93%. When training and testing against the entire data set, the accuracy was, to two decimal places, 74.83%.

It can be seen that when using the training set as the testing set, the prediction accuracy is significantly higher. However, this result is misleading. When testing using the training set, the classifier is classifying a data set for which it has already been shown the solutions, so a higher accuracy is expected. The accuracy is misleading as it is not classifying anything new, it is simply classifying what it has already seen. Conversely, the lower accuracy for the holdout strategy can be expected as the classifier in this case has not seen the testing data.