Replication

To replicate the results reported all you need to do is run the script and just adjust the 'project' variable on line 64 each time to change the dataset.

You can print to the _NB.csv files either the average for each metric over 10 runs (which the table in my report is based) or you can print each value from each of the 10 runs for each of the metrics. Currently the code prints each value from each of the 10 runs for each metric, to print out the average over 10 runs you just need to change the parameters on line 195 onwards to say "final_accuracy", "final_precision" etc.
My _NB.csv files currently show the printed values for both ways for each dataset.

To replicate the mann-whitney u test results, you will have to uncomment out the block of code at the bottom and for each metric of each dataset, you will need to paste in an array of 10 values from the baseline code for that metric of that dataset, and compare this to the solution's array of 10 values for the same metric. So to test recall of Caffe, you need to get the array of 10 recall values from the caffe dataset of the baseline solution (lab 1), paste it into the code over the variable "baseline_scores" and change the solution_scores parameter to recalls. This will give a result of 0.43, just like my report says.